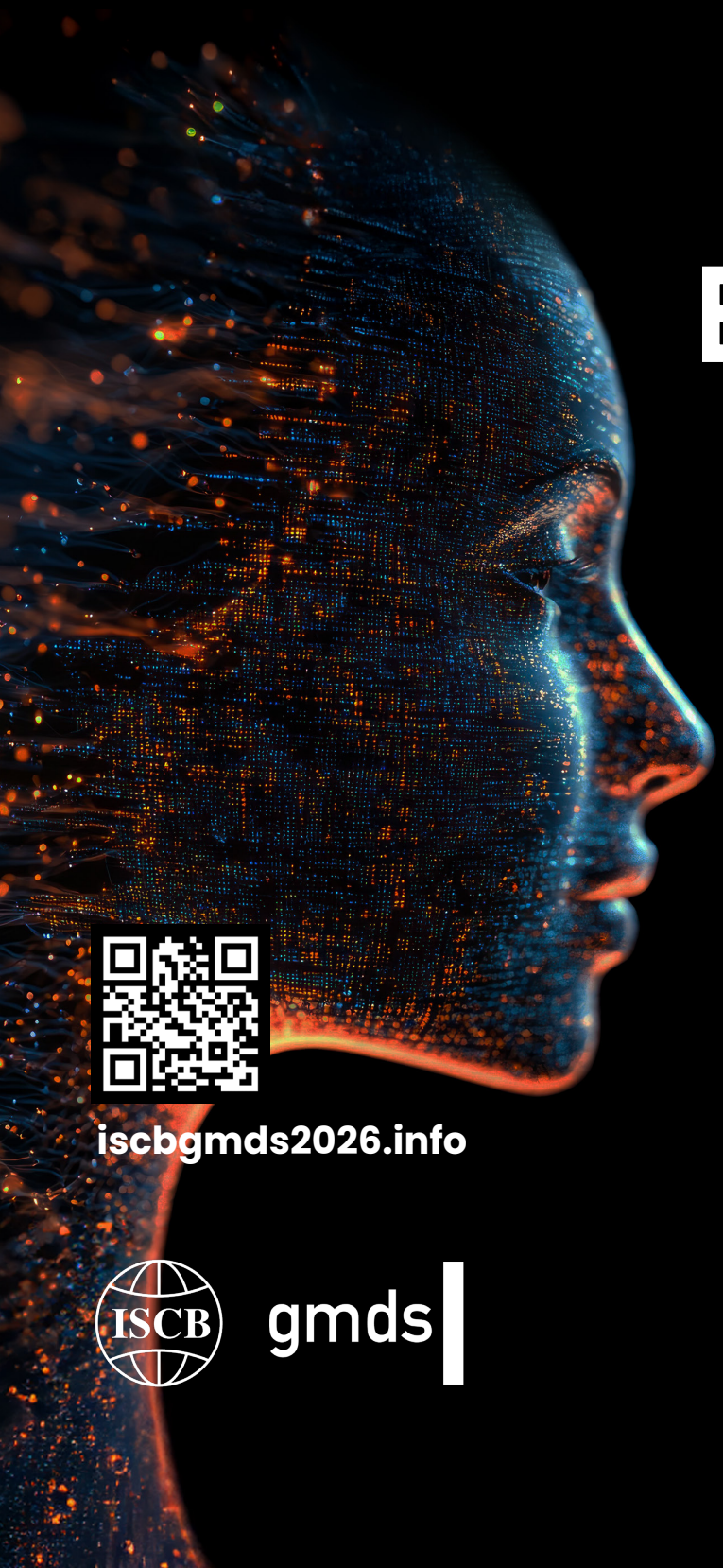
ISCB GMDS 2026

Conference Book



BIG DATA
small data
Your Data

Transitions in the age of biomedical AI

iscbgmds2026.info

FREIBURG | GERMANY

27.09. - 01.10.26

ISCB gmds

Contents

Welcome Messages **8**
Mayor of the City of Freiburg - Martin W. W. Horn 8
Conference President, SPC Chair - Nadine Binder 9
Conference President, LOC Chair - Harald Binder 10

About the Conference **11**
Conference Overview 11
Contact 13
Our Sponsors 14
Committees 15
Scientific program committee 15
Local organizing committee 17
Awards 18
ISCB Awards 18
GMDS Awards 20
Best Poster Awards 22

General On-Site Information **23**
Venue 23
Venue Floor Plans 25
Venue Access and Logistics 28
Practical Information 30
Conference App 31

Social Program **33**
Overview 33
Conference Excursions 36
Excursion to Freiburg's Local Mountain: Schauinsland 36
Kaiserstuhl Experience - Wine Tradition at site 36
Schlossberg Walk - Hike from the Old Town in the Vines 37
City Tour with Cathedral Tower Ascent 37
"Schwarzwald-Mädel" City Tour 38
Freiburg - City Rally 38
Traditional City Tour - A Journey Through Time 39
Wine Tasting at "Alte Wache" in the Old Town 39

Culinary City Tour - Chocolate 40
Cocktail Course - Typical Black Forest Cocktails 41

Conference Schedule **42**

Pre-Conference Courses **48**
Sun 27.09.2026 48
Performing Comparative Effectiveness Research Using Network Meta-Analysis: An Introduction to Theoretical and Practical Advances 48
Beyond "Randomized": Designing and Evaluating Randomization in Clinical Trials 50
Efficient Bayesian Analysis of Longitudinal and Survival Data using R-INLA 52
Causal inference for time-to-event outcomes - with practical applications in R 54
Prediction Under Interventions 56
Clinical Text Mining under Real-World Constraints 58
Genomic AI Agents for Personal Genomics: Hands-on Analysis with Just-DNA-Lite and LLM agents 60

Opening Session - Welcome to ISCB GMDS 2026 **62**

Keynote Sessions **63**
Cathie Sudlow: Future of Population Health Research: a UK Health Data Perspective 63
Marylyn Ritchie: The End of Data-Driven Medicine: Entering the Era of AI-Native Science 64
Annette Kopp-Schneider: Leveraging External Data in Clinical Trials: Principles, Perils & Potential 66
Andre Dekker: AI for Better Health Care 68

Invited Sessions **70**
Mon 28.09.2026 70
11:00–12:30 70
Biomedical Knowledge Integration 70
13:30–15:00 73
Causal inference with truncation by death and missing data: from estimands to estimation 73
15:30–17:00 76
Innovations in Master Protocols Evaluating Multiple Treatments and (Rare) Diseases 76
Tue 29.09.2026 79
11:00–12:30 79
AI/ML-Driven Causal Inference with Real-World Data Complexities 79
Wed 30.09.2026 82
09:00–10:30 82
Accelerating evaluation and adoption of complex digital outcome measures in trials 82
11:00–12:30 85
Biostatistics for personalized medicine 85
14:00–15:30 88

Evaluating Fairness in Clinical Prediction Models and AI: Current and Emerging Topics . 88
Signature Session: AI and Me 92

Workshops **93**
Mon 28.09.2026 93
11:00–12:30 93
Designing Next-Generation Respiratory Virus Trials: Estimands, Core Outcomes, and Adaptive Pandemic-Ready Frameworks 93
From Research Question to Access: Using Nationwide Claims Data Today - and Electronic Health Record Data Tomorrow 97
Developing and Executing Privacy-Preserving Federated Analyses with FLAME 99
From Theory to Practice in the EU HTA JCA: Early Insights and Lessons Learned 103
13:30–15:00 106
Analysis and Modelling of Infectious Diseases in the Age of AI: Lessons from COVID-19 106
Between Data, Practice, and Evidence: (Inter)Nationally EHR Research 111
Mapping metadata schemas to the HealthDCAT-AP standard: Lessons learned and hands on community examples (1/2) 113
A Guided Experiment on Clinical Trial Data Sharing Quality and Data Reuse (1/2) 117
15:30–17:00 120
Uncertainty in Clinical Predictions 120
Large Language Model Informed Real-World Cohort Data 123
Mapping metadata schemas to the HealthDCAT-AP standard: Lessons learned and hands on community examples (2/2) 126
A Guided Experiment on Clinical Trial Data Sharing Quality and Data Reuse (2/2) 128
Tue 29.09.2026 130
11:00–12:30 130
Reliable Subgroup Identification and Analysis - Results of GMDS Biostatistics Competition 2026 130
Introduction to Individual Participant Data Meta-Analysis 132
On the Ethical and Technical Aspects for Deployment and Evaluation of AI-Driven Clinical Decision Support Systems 133
13:30–15:00 139
Sponsor Workshop: SAS Data Maker + Friedrich Wingert Stiftung 139
Wed 30.09.2026 141
09:00–10:30 141
AI-driven Therapeutic Modelling in Biomedical Informatics (1/2) 141
Benchmarking AI scientists 143
Forum Digital Medicine and Artificial Intelligence (1/4) 146
Regulatory Topics for Statisticians: Implementation of Risk-Based Quality Management from Biostatistical Perspective 148
Ordinal Outcomes: Friend or Foe? 151
11:00–12:30 155
AI-driven Therapeutic Modelling in Biomedical Informatics (2/2) 155

Forum Digital Medicine and Artificial Intelligence (2/4) 158
Population-Adjusted Indirect Comparisons in the Framework of Health Technology Assessment 162
AI supported Data Retrieval and Data Management in the Health Domain 167
Use of Bayesian Methods for Clinical Trial Design and Analysis (1/2) 171
13:00–14:00 175
Forum Digital Medicine and Artificial Intelligence (4/4) 175
14:00–15:30 177
Forum Digital Medicine and Artificial Intelligence (3/4) 177
What Makes a Trial Complex? Evidence, Experience, and AI 180
Teaching Statistical Literacy in Times of Political Changes 183
Deep Dive Into eHealth With the Project MINT Meets Medicine 188
Use of Bayesian Methods for Clinical Trial Design and Analysis (2/2) 192

Oral Sessions **193**
Mon 28.09.2026 193
11:00–12:30 193
Frameworks for Clinical Research 193
Longitudinal and Joint Models of Clinical Trajectories 200
Design and Analysis Choices in Clinical Trials 207
Missing Data and Measurement Error 214
Digital Health Tools and Engagement 221
Semantic Health Data and Meta-Research 229
Survival Inference 236
13:30–15:00 242
LLMs for Clinical Workflows 242
Learning Latent Clinical Trajectories 250
Adaptive and Comparative Trial Design Strategies 258
Prediction and Imputation with Missing Data 264
Health Data Infrastructure and Secondary Data Use 271
Diagnostic Evidence Synthesis and Test Evaluation 278
Survival Prediction Models 285
15:30–17:00 293
Clinical Imaging 293
GMDS Awards 300
Adaptive Trial Inference and Error Control 308
Infectious Disease Modelling 315
Digital Health Education and Transformation 322
Network Meta-Analysis and Evidence Reliability 329
Dynamic Survival Modeling and Prediction 336
Tue 29.09.2026 343
11:00–12:30 343
LLMs for Clinical Text and Data 343
Clinical Trajectories and Events 351

Small Sample and Sequential Trial Methods ... 359
Robust Causal and Statistical Methods for Complex Data ... 366
Data Foundations for Clinical AI ... 374
Health Technology Assessment ... 382
Strategies for Prediction Models ... 389
13:30–15:00 ... 395
Data Quality, Selection, and Measurement ... 395
Statistical Learning under Uncertainty ... 403
Tailoring Statistical Models ... 410
Strategies for Evidence Synthesis ... 418
Adaptive Monitoring and Borrowing in Trials ... 425
Target Trial Emulation ... 432
Research-Ready Data Integration and Software Workflows ... 439
Prediction and Treatment ... 447
Treatment Effect Heterogeneity and Clinical Interpretation ... 454
Building and Updating Prediction Models ... 461
Wed 30.09.2026 ... 468
09:00–10:30 ... 468
Dose Finding and Adaptive Trial Design ... 468
Causal Inference with Survival and Competing Events ... 475
Interoperable Data and Digital Evaluation ... 483
Omics Discovery ... 491
Strategies in Clinical Trials ... 498
Recurrent Events Modelling ... 505
11:00–12:30 ... 512
Clinical Prediction ... 512
Basket Trials and Adaptive Borrowing ... 520
Causal Estimands and Biases ... 528
Federated Health Data Systems ... 535
Diagnostic Estimands and Modeling ... 543
Advances in Multi-State Modeling ... 550
14:00–15:30 ... 557
Platform and Cluster Trial Designs ... 557
Omics in Modeling ... 564
Causal Estimands and External Controls ... 572
Real-World Evidence, Registries, and Surveillance ... 579
Cancer Screening ... 587
Time-to-Event Methods and Meta-Analysis ... 595

Poster Sessions **603**
Mon 28.09.2026 ... 603
P1 - Digital Health and Data Systems ... 603
P2 - Prediction and Personalized Medicine ... 639
P3 - Clinical Trials ... 665

P4 - Causal Methods and Epidemiology 682
P5 - Time-to-event and Longitudinal Methods 699
P6 - Evidence and Decision Sciences 721
P7 - AI and Biomedical Data 739
Tue 29.09.2026 750
P1 - Digital Health and Data Systems 750
P2 - Prediction and Personalized Medicine 783
P3 - Clinical Trials 810
P4 - Causal Methods and Epidemiology 828
P5 - Time-to-event and Longitudinal Methods 847
P6 - Evidence and Decision Sciences 865
P7 - AI and Biomedical Data 882
Wed 30.09.2026 893
P1 - Digital Health and Data Systems 893
P2 - Prediction and Personalized Medicine 925
P3 - Clinical Trials 952
P4 - Causal Methods and Epidemiology 970
P5 - Time-to-event and Longitudinal Methods 988
P6 - Evidence and Decision Sciences 1009
P7 - AI and Biomedical Data 1027

Closing Session **1037**

Mini Symposia **1038**
SmallData Symposium 1038
SmallData Symposium Session 1: Similarity 1040
SmallData Symposium Session 2: Transfer 1042
SmallData Symposium Session 3: Uncertainty 1053
Poster sessions 1056
Poster Session: P1 - SmallData Similarity 1056
Poster Session: P2 - SmallData Transfer 1067
Poster Session: P3 - SmallData Uncertainty 1079
STRATOS Symposium 1087
STRATOS Symposium Session 1A: Methodological Aspects 1089
STRATOS Symposium Session 1B: Methodological Aspects 1091
STRATOS Symposium Session 2A: Practical Aspects of Medical Research 1094
STRATOS Symposium Session 2B: Practical Aspects of Medical Research 1096
Early Career Researcher Day 1100
Early Career Researcher Day Session 1 1102
Early Career Researcher Day Session 2 1103
Early Career Researcher Day Session 3 1104

SC/AG Meetings **1106**

Author Index 1111

Welcome Messages

Mayor of the City of Freiburg - Martin W. W. Horn

Source: Stadt Freiburg

Dear conference participants,

I am really delighted that the 47th Annual Conference of the International Society for Clinical Biostatistics together with the 71st Annual Conference of the German Association for Medical Informatics, Biometry and Epidemiology is being held in Freiburg. A warm welcome to all of you! Welcome to Freiburg!

This conference and the City of Freiburg: they're a perfect match. Firstly, because Freiburg is a leading centre for science and research. Freiburg is home to one of Germany's oldest and most prestigious universities, as well as several other institutions of higher education. With more than 15,000 employees, our university hospital is one of the largest in all of Germany. With five institutes and approximately 2,500 employees, Freiburg is the largest Fraunhofer location in Germany. Our city is a major hub for medical technology and pharmaceuticals - with nearly 40,000 people employed in this industry. And, Freiburg is one of the top 10 startup hubs in all of Germany - not least because startups here benefit from a very special environment of cutting-edge research and companies open to innovation.

And the second reason why this conference and the City of Freiburg are such a good fit: Because Freiburg is not only one of the sunniest cities in Germany, but also has a lot of sunshine in its heart. Our city stands for openness, diversity and hospitality.

Freiburg is a city of science and research, but it has much more to offer. I sincerely hope that you will find some time to discover our city in all its wonderful facets: It wasn't long ago that Freiburg even made it into Lonely Planet as a top travel destination. Why? Because of Freiburg's "enviable quality of life." Because bicycles are the number one mode of transportation in Freiburg. Because of our wonderful old town with its many cafes. And because we have the "Bachle" and the "Munster" Cathedral. Or in short: because Freiburg is one of the "most youthful, relaxed, and sustainable cities in the country".

Dear conference participants, I wish you many new insights and encounters, as well as unforgettable days in our city. Welcome to Freiburg!

Kind regards

Martin W. W. Horn

Mayor of the City of Freiburg

Conference President, SPC Chair - Nadine Binder

Dear colleagues,

A good scientific conference should do more than confirm what we already know. It should surprise us with an idea we had not considered, a question we had not thought to ask, or a conversation with someone we would otherwise never have met. With that hope, it is a great pleasure to welcome you to Freiburg for ISCB GMDS 2026.

In shaping the scientific program, our ambition was to create more than a meeting of the two scientific societies, ISCB and GMDS. We wanted to bring together communities that approach biomedical data from different perspectives, traditions and areas of expertise, because some of the most interesting things happen when different ways of thinking meet.

This feels particularly timely. Artificial intelligence is rapidly changing biomedical research, yet many fundamental questions remain: How do we learn reliably from complex or limited data? How do we quantify uncertainty? How do we design studies that answer the questions that matter? And how do we ensure that methodological innovation ultimately serves patients, researchers and society?

Our theme, "BIG DATA, small data, Your Data: Transitions in the age of biomedical AI," captures these contrasts: big data and small data, established methods and emerging ones, excitement about what is becoming possible and healthy skepticism about what it actually means. You will find these conversations throughout the program, in keynotes and invited sessions, contributed talks and posters, workshops, courses, mini-symposia and interactive formats.

A program like this is a collective achievement. I sincerely thank the Scientific Program Committee, the organizers of workshops and courses, and the editors and reviewers of the GMDS full papers. Most importantly, I thank everyone who submitted their research; your contributions are the foundation of this conference. My sincere thanks also go to the Local Organizing Committee and everyone working behind the scenes to make this meeting possible.

But the program in this book is only the framework. The conference itself begins when we start talking to one another. So be curious. Ask questions. Attend a session outside your usual field. Talk to someone you do not yet know. Share an unfinished idea. Challenge an assumption, including your own.

On behalf of the Scientific Program Committee, I wish you an inspiring conference, stimulating discussions, unexpected connections, and an enjoyable stay in Freiburg.

Welcome to ISCB GMDS 2026!

Nadine Binder

Chair, Scientific Program Committee

Conference President, LOC Chair - Harald Binder

Welcome to ISCB GMDS 2026!

Welcome to interdisciplinarity: to biostatistics, medical informatics, bioinformatics, epidemiology, and the many areas where these disciplines meet, overlap, and inspire each other.

Welcome to experiencing what our disciplines can contribute in a rapidly changing world, as biomedical AI, new forms of data, and new methodological possibilities challenge us to rethink established approaches. Welcome to Big Data, Small Data, Your Data, and to exploring the transitions that lie ahead.

Welcome to a showcase of the Collaborative Research Center Small Data, illustrating how interdisciplinary collaboration can make AI happen in setting that seemed too challenging. And welcome to Freiburg, a city shaped by science and scholarship and home to one of Germany's oldest universities.

Most importantly, welcome to discussions with peers: to exchanging ideas, questioning assumptions, making unexpected connections, and sparking change that matters. Welcome to five intense days of talks, posters, workshops, courses, encounters, and conversations. Welcome to food for thought, and to thoughts that may travel well beyond these five days.

This conference program and abstract book of offers a first glimpse of what awaits you. I hope that discovering its contents will raise your curiosity, lead you towards topics beyond your usual field, and provide starting points for many fruitful conversations.

Welcome also to the shaping of what will hopefully become some great memories.

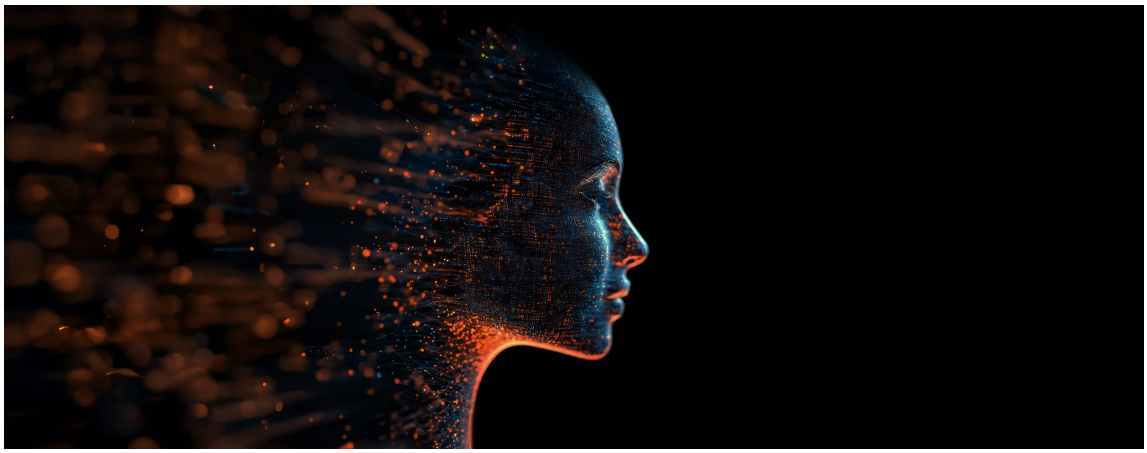
Welcome to the future.

Harald Binder

Chair, Local Organizing Committee

About the Conference

Conference Overview



ISCB GMDS 2026

ISCB

GMDS

Come with us and change the world

We are looking forward to jointly hosting the 47th Annual Conference of the International Society for Clinical Biostatistics (ISCB) together with the 71st Annual Conference of the German Association for Medical Informatics, Biometry and Epidemiology (GMDS) from September 27 to October 1, 2026 in Freiburg im Breisgau, Germany.

The rise of AI techniques and the increased public awareness of the value of data present great opportunities to data science disciplines. In particular, medical data sciences, as represented by the GMDS, and application fields such as novel treatment paradigms in medicine, which are hotly debated at ISCB, promise great impact of methodological developments, including transitions in research paradigms. This can optimally be achieved jointly with other disciplines, in particular mathematics, computer science, and also close collaboration with biomedical researchers, as exemplary reflected in the Freiburg CRC Small Data.

The ISCB GMDS 2026 Conference, together with the CRC Small Data Symposium, will provide a scientific platform for exchange on these exciting novel developments and discussing how our disciplines master the corresponding transitions.

Conference presidents

Harald Binder
LOC Chair

"I am grateful for the opportunity to host the conference as an interdisciplinary methods platform that will propel advancements in modeling and AI."

Nadine Binder
SPC Chair

"It is exciting to design a highly interdisciplinary program, reflecting the joint contributions of all methods disciplines, and I am looking forward to many contributions from the scientific communities."

About Freiburg

Freiburg is a scenic, vibrant university city and home to one of the oldest universities in Germany. Founded in 1120, picturesque Freiburg is located in southwest Germany's black forest. Locally rooted yet well connected at the European and global levels. Freiburg enjoys an excellent reputation as a science location with a focus on medicine, sustainability and innovation as well as education, science and research. The conference venue is situated in the heart of the city. Freiburg's Konzerthaus can look back on a history of over 10.000 hosted events and enjoys a high reputation far beyond the region.

Contact

1. Emergency & Medical Help

112: Emergency - ambulance / fire brigade

110: Police

116 117: Medical on-call service for urgent, non-life-threatening medical problems

0761 19240: Poison Control Centre Freiburg

0800 00 22 8 33: Emergency pharmacy service

116 116: Central service for blocking cards and some electronic banking services

2. Essential Local Information

International country code: +49

Freiburg area code: 0761; from abroad: +49 761

(When calling Germany from abroad, omit the initial 0 of the area/mobile code.)

3. ISCB GMDS 2026 Contacts

Conference Office & Organiser

Registration & logistics: iscbgmds2026@intercongress.de

Scientific content & formats: bemb.iscbgmds2026@uniklinik-freiburg.de

Submission-system technical support: helpdesk@webtek.at

During the Conference

Info Desk: General enquiries, lost & found and on-site guidance

Conference App: Personal schedule, programme updates and notifications

On-site emergency: Call 112 first in a life-threatening emergency; inform the conference team when safe

Our Sponsors

We sincerely thank all sponsors and partners for their valuable support of the ISCB GMDS 2026 conference.

Institutional & Platinum Sponsors

Silver Sponsors

Bronze Sponsors & Media Partners

Committees

Scientific program committee

Nadine Binder (SPC Chair)

Harald Binder

Adriani Nikolakopoulou

Andre Karch

Andre Scherag

Antonia Zapf

Cecile Proust-Lima

Dagmar Waltemath

Diane Uschner

Dominic Magirr

Fabian Prasser

Frank Ueckert

James Carpenter

Jan Gorodkin

Klaus Jung

Laure Wynants

Liangyuan Hu

Margarita Moreno-Betancur

Maja Pohar Perme

Maria Grazia Valsecchi

Maria Kalweit

Martin Posch

Nancy Leung

Nicole Radde

Spiros Denaxas

Suzanne Cadarette

Stefan Michiels

Stijn Vansteelandt

Ulrich Sax

Valeria Vitelli

Local organizing committee

Harald Binder (LOC Chair)

Nadine Binder

Sania Simran (Conference Manager)

Marie Weigold

Lea Grober

Olivia Kuhnrich

Claudia Giuliani

Derek Hazard

Georg Koch

Jean-Pierre Gnimatin

Jorg Sahlmann

Ksenija Fahrlander

Lukas Maximilian Horstmeier

Max Behrens

Moritz Hess

Olga Sieber

Sophia Nolde

Awards

ISCB Awards

Student Conference Award

The ISCB Student Conference Awards (StCA) recognize outstanding contributions by student researchers to the scientific programme of the conference. The awardees were selected by the ISCB Award Committee through a review process.

We warmly congratulate this year's StCA awardees on their selection and their outstanding contributions to ISCB GMDS 2026.

The Awardees

Nasrin Sultana: A Copula-Based Semi-Competing Risks Model for Recurrent and Terminal Events in Presence of Cure Fraction and its Efficient Estimation via Stochastic Optimization.

Presented in: Recurrent Events Modelling – Wednesday, 30 September 2026, 10:00–10:15, K2.

Arnau Garcia-Fernandez: A Strictly Proper Metric to Evaluate Dynamic Predictions for Left-Truncated and Right-Censored Data with Asymmetric Misclassification.

Presented in: Dynamic Survival Modeling and Prediction – Monday, 28 September 2026, 16:45–17:00, K2.

Nana-adjoa Kwarteng: Assessing the Impact of Model Assumptions in Network Meta-Regression: A Simulation Study.

Presented in: Network Meta-Analysis and Evidence Reliability – Monday, 28 September 2026, 15:45–16:00, N4.

Zhi Cao: Estimating Subtrial-specific Optimal Biological Doses in Basket Trials Permitting Robust Borrowing of Information.

Presented in: Basket Trials and Adaptive Borrowing – Wednesday, 30 September 2026, 11:45–12:00, K5.

Timothy Osborne: Identifying Harmful Drug-Drug Interactions Using Bayesian Profile Regression and Stochastic Variational Inference Applied to Population-Scale Linked Routine Health Data in Wales.

Presented in: Robust Causal and Statistical Methods for Complex Data – Tuesday, 29 September 2026, 12:15–12:30, N1.

Jonatan Hedberg: Non-parametric Estimation of Excess Recurrent Events.

Presented in: Recurrent Events Modelling – Wednesday, 30 September 2026, 09:30–09:45, K2.

Conference Attendance Support Program (CASP)

The following awardees were selected for the Conference Attendance Support Program (CASP):

- Maliha Binte Alauddin

- Safia Awan
- Lun Hu
- Md. Monowar Hossain
- Md. Hasibul Islam Jitu
- Moss. Maharun Nahar Jui
- Hamdi Furkan Kepenek
- Md Hasinur Rahaman Khan
- Anita Pal
- Daniel J Phillips
- Abdur Rasheed
- Sandra Rjeschni
- Roshni Tony
- Ebru Turgal
- Kim Zhipei Wang
- Hui Wen
- Jamie Wong

GMDS Awards

GMDS Förderpreise

The GMDS Förderpreise recognize outstanding scientific work across the fields represented by the German Association for Medical Informatics, Biometry and Epidemiology (GMDS).

We warmly congratulate this year's awardees on their achievements.

The Awardees

Henrike Solveen: A Simulation-Based Comparison of Winner's Curse Correction Methods in the Context of Multi-Group Animal Experiments.

Presented in: GMDS Awards – Mon, 28.09.2026, 15:35–15:40, K3.

Selin Gürbulak: Introduction of the Electronic Health Professional Card (eHBA) at a University Hospital, Using Hannover Medical School (MHH) as an Example - Development of an Ideal Process.

Presented in: GMDS Awards – Mon, 28.09.2026, 15:40–15:45, K3.

Angelika Waßmuth: The Effect of Smoking on the Composition of the Anterior Nares' Microbiome.

Presented in: GMDS Awards – Mon, 28.09.2026, 15:45–15:50, K3.

Anna Linger: Sample Size Recalculation in a Hybrid Adaptive Clinical Trial Design for Time-to-Event Endpoints.

Presented in: GMDS Awards – Mon, 28.09.2026, 15:55–16:00, K3.

Celine Magnus: Design and Development of an Assistant-Supported Validation Process for EDC Systems as Software-as-a-Service in Clinical Research.

Presented in: GMDS Awards – Mon, 28.09.2026, 16:00–16:05, K3.

Malte Mensching-Buhr: Unsupervised Learning of Cell Type Distributions in Spatial Transcriptomics Data via Spatially-Informed Non-Negative Matrix Factorization.

Presented in: GMDS Awards – Mon, 28.09.2026, 16:05–16:10, K3.

Manuel Hecht: Investigator-Initiated versus Industry-Sponsored Trials - Visibility and Relevance of Randomized Controlled Trials in Clinical Practice Guidelines (IMPACT).

Presented in: GMDS Awards – Mon, 28.09.2026, 16:15–16:20, K3.

Najia Ahmadi: GenRare CDM: A Standardized and Reusable Framework for Integrating Genomic, Phenotypic, and Clinical Rare Disease Data for Multi-Site AI-Driven Research.

Presented in: GMDS Awards – Mon, 28.09.2026, 16:20–16:25, K3.

Maren Hackenberg: Capturing Dynamics under Misspecification and Potential Transformation.

Presented in: GMDS Awards – Mon, 28.09.2026, 16:25–16:30, K3.

Paul Martini Preis

Carolin Herrmann: Adaptive Study Designs with Restricted Mean Survival Times and Sample Size Adaptations.

Presented in: GMDS Awards – Mon, 28.09.2026, 16:35–16:50, K3.

Best Poster Awards

The three best posters of the ISCB GMDS 2026 will be awarded during the closing session of the conference.

We are counting on the conference participants for selecting the awardees: each participant has been assigned three posters to evaluate, one per conference day.

Therefore, we encourage all participants to engage with poster presenters and evaluate the assigned posters.

The posters are evaluated based on these criteria:

- Relevance and scientific content (50%): How would you rate the presented science?
- Design and visual communication (20%): Is the work presented in a way which eases the understanding and grasping of essential results?
- Engagement (20%): Was the presenter motivated and engaged to explain the work and answer questions?
- Overall impression (10%): What is your overall impression of the quality of the poster?

For each category, a score from 1 to 10 can be assigned, where 1 = extremely poor and 10 = exceptional.

You can find the list of posters assigned to you by entering the email address with which you registered to the conference on this webpage: `https://iscbgmds2026.info/poster-awards-voting/`.

Disclaimer: poster assignments cannot be changed. We appreciate your efforts in evaluating the posters you can, and understand that you may be unable to evaluate all of them.

General On-Site Information

Venue

Konzerthaus Freiburg

Main conference venue

Konrad-Adenauer-Platz 1
79098 Freiburg im Breisgau
Germany

Novotel Freiburg am Konzerthaus

Konrad-Adenauer-Platz 2
79098 Freiburg im Breisgau
Germany

Welcome Reception

Konzerthaus Freiburg
Konrad-Adenauer-Platz 1
79098 Freiburg im Breisgau
Germany

Early Career Researcher Gathering

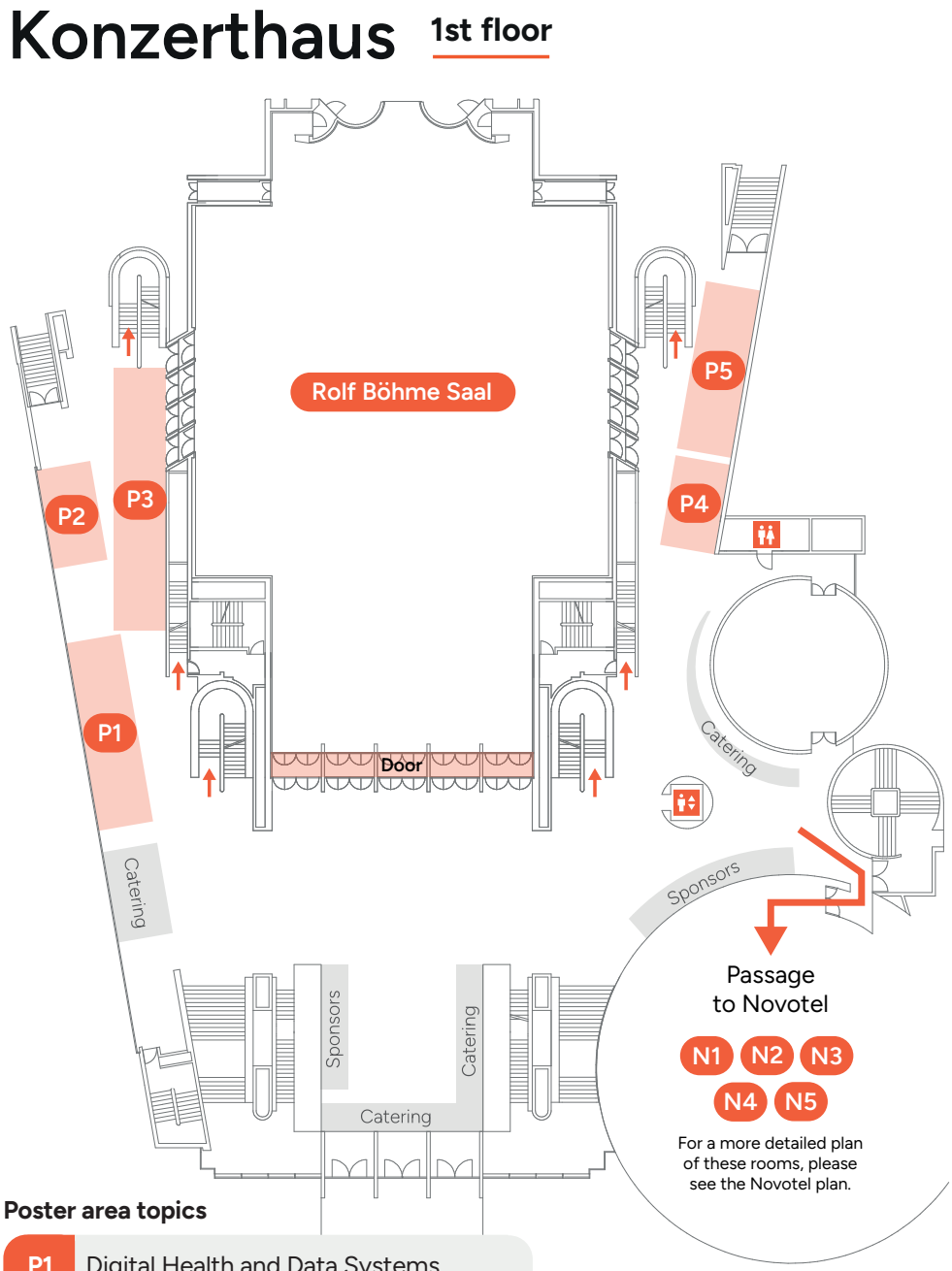
Feierling (Insel)
Gerberau 46
79098 Freiburg im Breisgau
Germany

Conference Dinner

Kultur & Bürgerhaus | Delcanto

Stuttgarter Str. 30
79211 Denzlingen
Germany

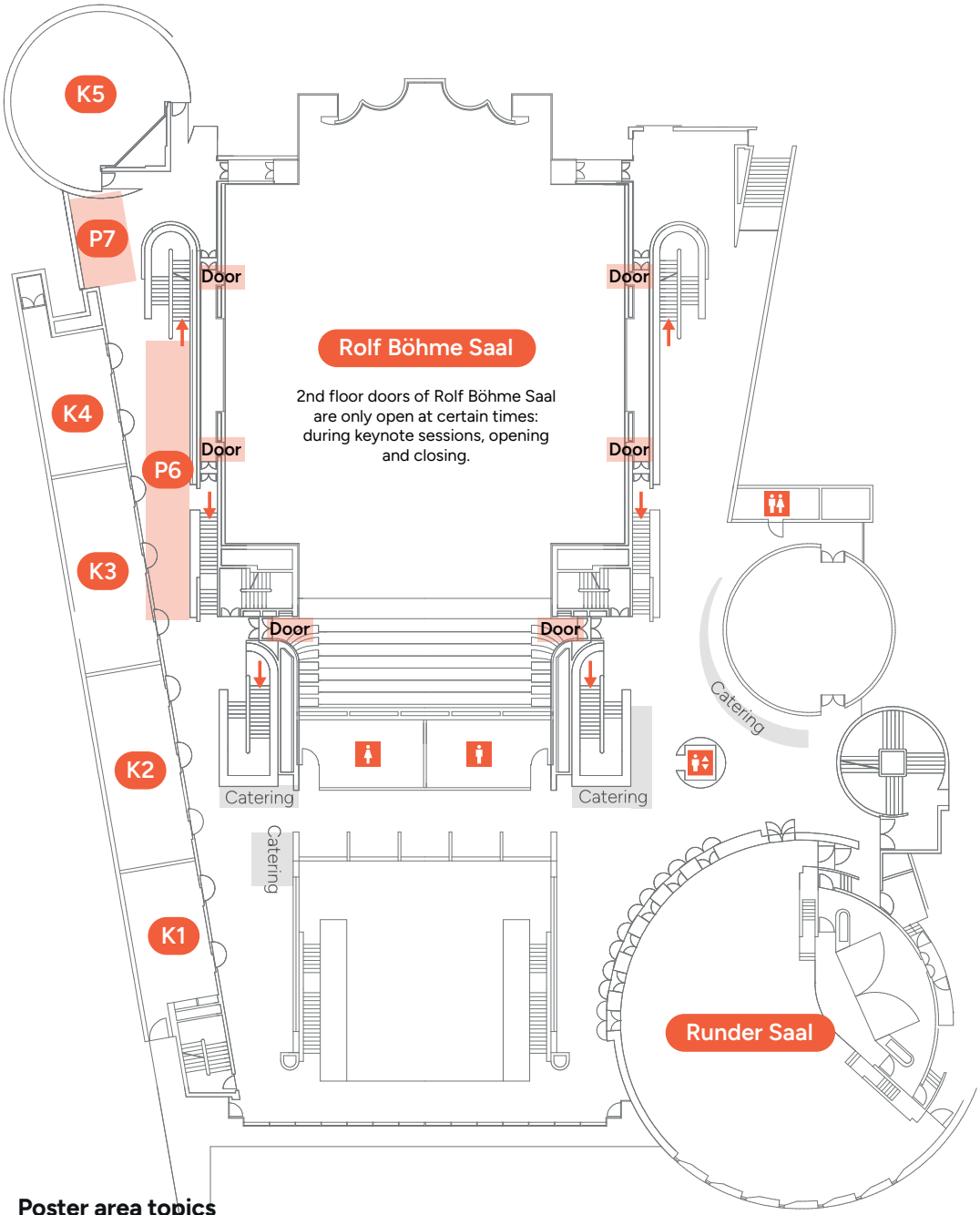
Venue Floor Plans



Poster area topics

P1	Digital Health and Data Systems
P2	Prediction and Personalized Medicine
P3	Clinical Trials
P4	Causal Methods and Epidemiology
P5	Survival and Longitudinal Methods

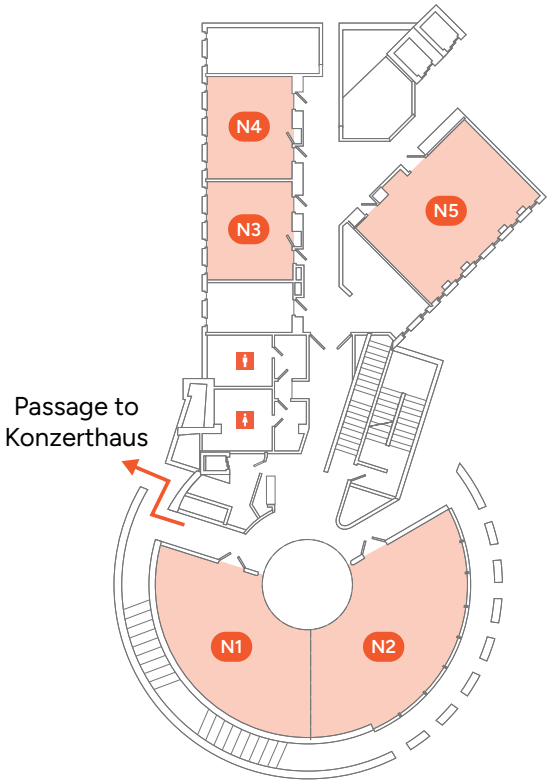
Konzerthaus 2nd floor



Poster area topics

P6	Evidence and Decision Sciences
P7	AI and Biomedical Data

Novotel



Venue Access and Logistics

Conference Office

The Conference Office is located in the entrance area of Konzerthaus Freiburg and will be open daily from Sunday, September 27 to Thursday, October 1, starting at 8:00 until the end of each conference day. They will offer assistance with inquiries, lost and found, and general guidance.

QR Code

Your personal QR code will be sent to you by email approximately one week before the conference starts. You will find it on your eTicket and in the wallet file included with your registration documents.

Please have your QR code ready on arrival. You will need it to:

- enter the conference venue - your QR code will be scanned at the entrance
- print your name badge at the Conference Office

Name Badge

After entering the venue, you can print your name badge at the Conference Office using your personal QR code.

Your name badge is your conference ID. Please wear it visibly at all times, as it is required to access the conference venue and all activities included in your registration, such as Pre-Conference Courses, Mini Symposia and social events.

Badge printing hours

Sunday, 27.09.2026: 08:00-17:15 CET (upper floor accessible to pre-conference course participants only)

Monday, 28.09.2026: 08:00-16:30 CET

Tuesday, 29.09.2026: 08:00-15:00 CET

Wednesday, 30.09.2026: 08:00-17:00 CET

Thursday, 01.10.2026: 08:00-16:30 CET

Conference Venue and Novotel Rooms

The conference takes place throughout the Konzerthaus Freiburg as well as in selected rooms within the Novotel. These Novotel conference rooms are directly accessible from inside the Konzerthaus. There is no need to leave the conference venue or use a separate entrance.

Please note that the only entrance to the conference is through the Konzerthaus. Upon arrival, please enter via the Konzerthaus and follow the on-site signage to your session room.

Please check the Conference App for the room assigned to your session.

Wi-Fi Access

Complimentary Wi-Fi will be available throughout the conference. Please note that there are two separate Wi-Fi networks: one for the Konzerthaus and one for the Novotel meeting rooms.

The network names and access details will be provided onsite and in your conference materials.

Lost and Found

If you have lost a personal item, please check with the Conference Office. All found items will be handed in and kept there.

If you find a lost item, please hand it in at the Conference Office.

Luggage Storage

A complimentary luggage storage area will be available at the Konzerthaus.

As storage space is limited, we kindly ask you to leave your luggage at your hotel whenever convenient.

Practical Information

Credit Cards

Credit cards are widely accepted. MasterCard & Visa dominate within Germany. American Express is accepted in some stores and of course you can use it with the ATM. Furthermore, you can check shop windows and entrance doors for credit card logos to see which cards you can use.

Electricity / Voltage

Electrical sockets in Germany are one of the two European standard electrical socket types: The "Type C" Europlug and the "Type E" and "Type F" Schuko. The electric current in Germany is 220 volts. Be aware that you may need a travel adapter for certain electric devices.

Languages

National language: German (Alemannic dialect alongside)

Many locals and business and service employees speak English, particularly the younger ones.

Time Zones

The time zone in Germany is Central European Time (UTC+01:00) and Central European Summer Time (UTC+02:00).

Telecommunications

Pay phones are no longer common in Germany and hard to find. The International Direct Dial code for Germany is +49 and the city code for Freiburg is 0761.

Useful Telephone Numbers:

Police 110

Medical emergency / accident services 112

Fire Service 112

Poison Control Centre Freiburg 0761 19240

Currency

The German monetary unit is the Euro (EUR).

You can get cash from any Geldautomat (ATM) at major train stations, airports, large shopping centers and in and outside banks. Within the Eurozone charges for ATM withdrawals and credit card payments include domestic payments and it is possible that you get domestic fees charged.

Gratuities / Tipping

In restaurants in Germany it is common to round up to the next Euro or add a few Euros (5 - 10 % for good service). Generally, you give the gratuity directly to the waiter or you add it to the bill. Often the waiter only tells you how much you owe. Then, you can tell how much you intend to pay when handing over the money. To leave the tip on the table is not common in Germany.

Conference App

ISCB GMDS 2026 Conference App, which will serve as your central companion throughout the conference. The app provides access to the detailed scientific program, important on-site information, and allows you to create your own personalized conference schedule. It will keep you informed with the latest updates and any changes to the conference program in real time.

Download the conference app "ISCB GMDS 2026"

Please select the appropriate app store for your device:

- [Apple App Store]
- [Google Play Store]

Login for accessing the detailed program

To view the conference program, you will need to log in to the app or the web version using your login credentials from our online submission system. Please note that these may not be the same as your conference registration platform credentials.

If you do not yet have an account in our submission system, simply select "Create new account" on the login page and follow the registration process.

Web version of the conference program

If you are unable to access the conference app, you can also view the conference program through our online submission system using a web browser on your preferred device. Simply log in to the online submission system, navigate to Start, and select Conference Program.

iOS App Store

iPhone / iPad

Google Play

Android

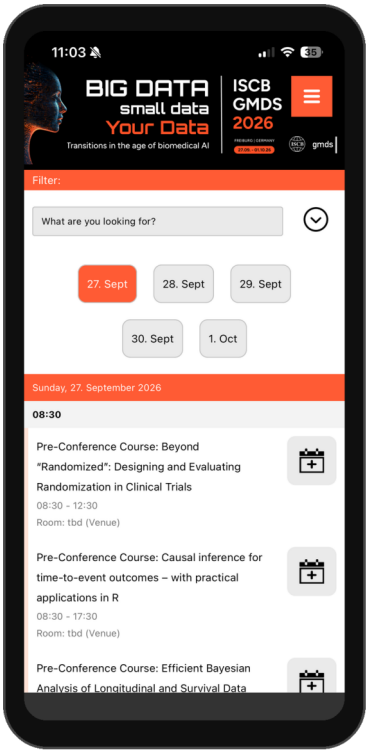
Create your personal schedule

One of the key features of the conference app is the ability to create your own personal schedule.

Simply browse the scientific program and add sessions to your personal schedule by clicking the calendar icon with the "+" symbol next to a session. All selected sessions will then be displayed in your personal schedule, making it easy to plan your conference experience and quickly find the presentations and events you would like to attend.

Your personal schedule is not binding and does not constitute a registration for individual sessions. However, it provides us with valuable information for room planning and helps us optimize the on-site experience for all participants.

The summer break is a great opportunity to explore the exciting scientific program and start putting together your own schedule-there is plenty of interesting scientific content to discover!



Social Program

Overview

The social program of ISCB GMDS 2026 offers a welcoming opportunity to continue conversations beyond the scientific sessions, discover Freiburg, and connect with colleagues in a more informal setting. From shared social events during the conference evenings to the conference excursions on Tuesday, September 29, 2026, the program is designed to encourage exchange, new contacts, and memorable experiences throughout the meeting.

All social events can be booked during conference registration.

Conference Evenings

Early Career Researcher Gathering 34
Welcome Reception 34
Conference Dinner 34

Conference Excursions

Excursion to Freiburg's Local Mountain: Schauinsland 36
Kaiserstuhl Experience - Wine Tradition at site 36
Schlossberg Walk - Hike from the Old Town in the Vines 37
City Tour with Cathedral Tower Ascent 37
"Schwarzwald-Mädel" City Tour 38
Freiburg - City Rally 38
Traditional City Tour - A Journey Through Time 39
Wine Tasting at "Alte Wache" in the Old Town 39
Culinary City Tour - Chocolate 40
Cocktail Course - Typical Black Forest Cocktails 41

Early Career Researcher Gathering

Start the conference week in a dedicated event area at Feierling (Insel), a modern space with wood and industrial design that is ideal for larger groups. Connect with fellow early career researchers and enjoy the same Baden inspired food, drinks and beer specialties as in the main Feierling.

Feierling (Insel), reserved event room, Gerberau 46, 79098 Freiburg im Breisgau, Germany. 27 September 2026, 18:00 - 22:00 CET.

Welcome Reception

Meet colleagues, speakers and organizers at the official Welcome Reception on the 28 September 2026. A perfect opportunity to start networking at the first main conference evening.

Konzerthaus, Konrad-Adenauer-Platz 1, 79098 Freiburg im Breisgau, Germany. 28 September 2026, 18:30 - 20:30 CET.

Conference Dinner

Enjoy an unforgettable conference evening at the Kultur & Bürgerhaus (Delcanto) in Denzlingen. The Conference Dinner brings everyone together for great food, lively conversations and a festive atmosphere.

Shuttle service available: departure from Konzerthaus Freiburg at 18:30 CET; return from 22:15 CET.

Kultur & Bürgerhaus, Stuttgarter Str. 30, 79211 Denzlingen, Germany. 30 September 2026, 19:00 - End.

Conference Excursions

Excursion to Freiburg's Local Mountain: Schauinsland

A journey to Freiburg's local mountain and guided tour through a mining museum.

Key Information

- Start: 29.09.2026 - 15:15 CET
- Duration: 3,5 hours
- Meeting Point: Pickup by a guide at the Konzerthaus Freiburg
- Transfer: Travel to the valley station of the cable car "Schauinslandbahn", ascent via cable car and return transfer to the Konzerthaus
- Accessibility: The 400m tour path from the mountain station to the mine entry is level and easy to walk.
- Requirements: No ladders or age restrictions. Street clothing is fine, though a jacket is highly recommended (it stays cool underground). Standard footwear is sufficient.

Experience the ultimate contrast of breathtaking views and subterranean history on this excursion to Schauinsland, Freiburg's iconic local mountain.

The journey begins with a ride to the valley station, where you'll board Germany's longest loop cable car. As you glide toward the summit, enjoy fascinating views of the city of Freiburg and the Rhine Valley, stretching all the way to the blue ribbon of the Vosges mountains.

The adventure continues underground with a tour through the museum mine. Enjoy an expert-led tour through the upper level of the visitor mine, including live machine demonstrations (compressed-air rock drill and overhead loader). Learn about the history, geology, and technology of the region. The tour includes a view into the 50-meter-deep transport shaft featuring mine locomotives and a technical collection.

After returning to the surface, we will bring you back to the city, having experienced the very best of Freiburg's "house mountain".

Kaiserstuhl Experience - Wine Tradition at site

Unveil a landscape shaped by fire and time on a journey through one of Europe's most unique viticultural treasures.

Key Information

- Start: 29.09.2026 - 15:15 CET
- Duration: 3,5 hours

- Meeting Point: Pickup by a guide at the Konzerthaus Freiburg
- Transfer: Travel to the Kaiserstuhl village Ihringen and back to the city centre by public transport
- Requirements & Accessibility: Paths are well-maintained, there are some steady inclines. Standard footwear is sufficient.

Rising majestically from the Rhine Valley, the Kaiserstuhl is a volcanic paradise where sun-drenched vineyards and Mediterranean flair create Germany's warmest wine region.

Discover the magic of the region on an immersive wine trail through its diverse landscape. Explore the evolution of viticulture from its ancient roots to modern-day techniques. From intriguing tales of geology to timeless traditions, come get inspired and experience the flavors and beauty of the Kaiserstuhl with all your senses!

Schlossberg Walk - Hike from the Old Town in the Vines

A perfect blend of urban charm and vineyard beauty.

Key Information

- Start: 29.09.2026 - 15:15 CET
- Duration: 3 hours
- Meeting Point: Pickup by a guide at the Konzerthaus Freiburg
- Transfer: Walk through the city to the Schlossberg
- Requirements & Accessibility: Paths are well-maintained, there are some steady inclines and stairs. Comfortable sneakers with decent grip are sufficient.

Perched directly above the historic Old Town, the Freiburg Schlossberg is the city's vibrant green heart and a true paradise for nature lovers and wine enthusiasts alike. This iconic hill blends rich history with lush nature, serving as the perfect gateway to experience the authentic spirit of the Black Forest foothills.

The wine hike across the Schlossberg offers more than just world-class wine; it treats you to breathtaking vistas over the city and the idyllic Kaiserstuhl landscape, stretching all the way to the Vosges Mountains.

The tour concludes in one of Germany's premier vineyards - right in the heart of the vines - where we'll savor two additional local wines paired with delicious regional snacks.

City Tour with Cathedral Tower Ascent

Guided walk through the Old Town and a climb up the "most beautiful tower on earth".

Key Information

- Start: 29.09.2026 - 15:15 CET
- Total Duration: Approx. 2 hours
- Meeting Point: Pickup by a guide at the Konzerthaus Freiburg
- Procedure: Guided tour through the lower Old Town leading to the Cathedral (Münster), climbing the Cathedral Tower and the tour ends at the Cathedral exit, followed by free time for participants

The undisputed highlight of this tour is the breathtaking view from the Cathedral Tower over Freiburg's Old Town and far into the impressive surrounding landscape. From the tower gallery, at the transition to the world-famous open-work spire, you can see across the Kaiserstuhl and the Rhine Valley all the way to the Vosges Mountains in France. On the other side, you can overlook the Dreisam Valley and the Black Forest.

But it isn't just the view that inspires visitors. During the ascent of 209 steps to the watchman's cabin (Türmerstube), you will experience the construction of the massive fir-wood bell frame. Started in 1290, it remains 80% original. After another 56 steps, you reach the tower platform. For those who want to go even higher, the gallery awaits at a height of 70 meters. High, higher - the very top!

"Schwarzwald-Mädel" City Tour

Key Information

- Start: 29.09.2026 - 15:15 CET
- Total Duration: Approx. 2 hours
- Meeting Point: Pickup by a guide in costume at the Konzerthaus Freiburg
- The Tour: A guided exploration of the city

Follow the charming Black Forest local girl "Schwarzwald Mädel", Bärbel, through the winding alleys of Freiburg, the gateway to the Black Forest. Let yourself be swept away by the vibrant history of both the city and the region. Discover fascinating facts and amusing anecdotes on this extraordinary walking tour through Freiburg's historic Old Town.

Freiburg - City Rally

A scavenger hunt through the enchanting city of Freiburg - above and below ground...

Key Information

- Start: 29.09.2026 - 15:15 CET
- Total Duration: Approx. 2 hours
- Meeting Point: Pickup at the Konzerthaus Freiburg
- The Tour: An exciting tour through the city centre with interactive challenges.

Experience Freiburg's unique flair through an exciting blend of culture, navigation challenges and activities. Perfect for those who want to discover Freiburg at their own pace, this interactive adventure invites you to explore the streets armed with a quest kit and special gear.

Forget dry facts - this mission is all about creativity, wit, and ingenuity. It's an engaging, self-guided journey that turns the historic center into your personal playground - above and below ground...

Traditional City Tour - A Journey Through Time

Key Information

- Start: 29.09.2026 - 15:15 CET
- Total Duration: Approx. 2 hours
- Meeting Point: Pickup by a guide at the Konzerthaus Freiburg
- The Tour: A guided tour through the city

Experience centuries of captivating history brought to life with a perfect blend of wit and expert knowledge. Our journey takes you along the iconic Freiburger Bächle - the city's famous splashing runnels - and through picturesque, cobblestone "Gässle" that whisper stories of the past.

From the city's most stunning squares to the ornate facades of historic merchant houses, we make our way toward the breathtaking Freiburg Minster, a masterpiece of Gothic architecture. But it's not just about the past: you'll also get an authentic glimpse into the vibrant, modern pulse of Germany's sunniest city.

Wine Tasting at "Alte Wache" in the Old Town

Key Information

- Start: 29.09.2026 - 15:15 CET
- Duration: Approx. 2 hours
- Meeting Point: Pickup by a guide at the Konzerthaus Freiburg
- The Experience: A walk through the city to the Alte Wache directly besides the Minster of Freiburg. Professional wine tasting with 5 wines.

The Alte Wache in Freiburg - the "gateway to Baden wine" It doesn't matter if you are from Freiburg, a wine connoisseur from the local region or a guest from further afield: thanks to its variety of wines and a stunning location on Münsterplatz square in central Freiburg, the Alte Wache draws in lots of visitors each year. Since it first opened back in 1997, it has become an institution for Baden wines and a popular place to meet in the centre of Freiburg.

During the guided wine tastings, the wine consultants and sommeliers take plenty of time to share the stories behind the enjoyment of Baden wines. Of course, a good sip is a must.

Culinary City Tour - Chocolate

Key Information

- Start: 29.09.2026 - 15:15 CET
- Duration: Approx. 2 hours
- Meeting Point: Pickup by a guide at the Konzerthaus Freiburg

Freiburg's past is woven with the elegance of four hundred years of Habsburg rule, a golden era that left behind a legacy far richer than just architecture. On this leisurely afternoon stroll, we celebrate the sophisticated Freiburg's coffeehouse tradition that continues to pulse through the city today.

While we like to keep a few secrets up our sleeve to ensure a sense of wonder, you can look forward to:

- A Journey Through Time: Discover how imperial history shaped the local flavors and social customs of the Black Forest capital.
- Hidden Gems: Wander past iconic landmarks and duck into charming corners that many travelers overlook.
- The Best of Freiburg: See the city from its most delicious perspective, where history and culinary delights meet.

Cocktail Course - Typical Black Forest Cocktails

Key Information

- Start: 29.09.2026 - 15:15
- Duration: Approx. 2 hours
- Meeting Point: Pickup by a guide at the Konzerthaus Freiburg

Welcome to the Black Forest Bar Academy! Whether you're a curious beginner or a home-bartender looking to level up, we are thrilled to have you here in the heart of Freiburg.

We'll start with a brief introduction to bar philosophy, where you'll get an overview of the glasses, equipment, and the basics of mixology. This will lay the groundwork before we pick up the shaker. Of course, during the theoretical part, you'll also learn about the basic spirits and cocktail families.

Then it's time to grab your shakers - ready, set, go!

ISCB GMDS 2026

Preliminary Program

BIG DATA small data Your Data

Transitions in the age of biomedical AI

Sunday, 27.09.2026

08:30 - 10:30	Parallel Pre-Conference Courses (Half- and Full-Day)
10:30 - 11:00	Coffee Break
11:00 - 12:30	Parallel Pre-Conference Courses (Half- and Full-Day)
12:30 - 13:30	Lunch Break
13:30 - 15:00	Parallel Pre-Conference Courses (Half- and Full-Day)
15:00 - 15:30	Coffee Break
15:30 - 17:30	Parallel Pre-Conference Courses (Half- and Full-Day)
18:00 - End	Early Career Researcher´s Gathering

Monday, 28.09.2026

09:00 - 10:30	Opening Session, Awards + Keynote Speaker 1 (45')
10:30 - 11:00	Coffee Break
11:00 - 12:30	Invited Session, Parallel Contributed Sessions and Workshops
12:30 - 13:30	Lunch Break SC & AG meetings
13:30 - 15:00	Invited Session, Parallel Contributed Sessions and Workshops
15:00 - 15:30	Coffee Break
15:30 - 17:00	Invited Session, Parallel Contributed Sessions and Workshops
17:00 - 18:30	GMDS Mitglieder-versammlung
18:30-20:30	Welcome Reception

Tuesday, 29.09.2026

09:00 - 10:30	Keynote Speakers 2 + 3 (45' each)
10:30 - 11:00	Coffee Break
11:00 - 12:30	Invited Session, Parallel Contributed Sessions and Workshops
12:30 - 13:30	Lunch Break SC & AG meetings
13:30 - 15:00	Parallel Contributed Sessions and Workshops
15:15 - End	Conference Excursions / Social Afternoon

Wednesday, 30.09.2026

09:00 - 10:30	Invited Session, Parallel Contributed Sessions and Workshops
10:30 - 11:00	Coffee Break
11:00 - 12:30	Invited Session, Parallel Contributed Sessions and Workshops
12:30 - 13:40	AGM ISCB
12:30 -14:00	Lunch Break SC & AG meetings
14:00 - 15:30	Invited Session, Parallel Contributed Sessions and Workshops
15:30 - 16:00	Coffee Break
16:00 - 17:30	Keynote Speaker 4 (45'), Poster Awards + Closing
19:00 - End	Conference Dinner

Thursday, 01.10.2026

09:00 - 10:30	SmallData Symposium, STRATOS Symposium, Early Career Researcher Day
10:30 - 11:00	Coffee Break
11:00 - 12:30	SmallData Symposium, STRATOS Symposium, Early Career Researcher Day
12:30 - 13:30	Lunch Break
13:30 - 15:00	SmallData Symposium, STRATOS Symposium, Early Career Researcher Day
15:00 - 15:30	Coffee Break
15:30 - 17:00	STRATOS Symposium

ISCB GMDS 2026

Preliminary Program (Day view)

BIG DATA small data Your Data
Transitions in the age of biomedical AI

Sunday, 27.09.2026					
	K1	K2	K3	K4	K5
08:30 - 10:30	Performing Comparative Effectiveness Research Using Network Meta-Analysis: An Introduction To Theoretical And Practical Advances	Beyond “Randomized”: Designing and Evaluating Randomization in Clinical Trials	Efficient Bayesian Analysis of Longitudinal and Survival Data using R-INLA		Causal Inference for Time-to-Event Outcomes – With Practical Applications in R
10:30 - 11:00	Coffee Break				
11:00 - 12:30	Performing Comparative Effectiveness Research Using Network Meta-Analysis: An Introduction To Theoretical And Practical Advances	Beyond “Randomized”: Designing and Evaluating Randomization in Clinical Trials	Efficient Bayesian Analysis of Longitudinal and Survival Data using R-INLA		Causal Inference for Time-to-Event Outcomes – With Practical Applications in R
12:30 - 13:30	Lunch Break				
13:30 - 15:00	Performing Comparative Effectiveness Research Using Network Meta-Analysis: An Introduction To Theoretical And Practical Advances	Prediction Under Interventions	Clinical Text Mining under Real-World Constraints	Genomic AI-Agents for Personal Genomics: Hands-On Analysis With Just-DNA-Lite and LLM Agents	Causal Inference for Time-to-Event Outcomes – With Practical Applications in R
15:00 - 15:30	Coffee Break				
15:30 - 17:30	Performing Comparative Effectiveness Research Using Network Meta-Analysis: An Introduction To Theoretical And Practical Advances	Prediction Under Interventions	Clinical Text Mining under Real-World Constraints	Genomic AI-Agents for Personal Genomics: Hands-On Analysis With Just-DNA-Lite and LLM Agents	Causal Inference for Time-to-Event Outcomes – With Practical Applications in R
18:00 - 22:00	Early Career Researcher´s Gathering				

ISCB GMDS 2026

Preliminary Program (Day view)

BIG DATA small data Your Data

Transitions in the age of biomedical AI

Monday, 28.09.2026													
	Rolf Böhme Saal	Runder Saal	K1	K2	K3	K4	K5	N1	N2	N3	N4	N5	P1-P7
09:00 - 09:45	Opening Session												Partner & Poster Exhibition
09:45 - 10:30	Keynote: Cathie Sudlow												
10:30 - 11:00	Coffee Break												
11:00 - 12:30	IS: Biomedical Knowledge Integration	Survival Inference	WS: Designing Next-Generation Respiratory Virus Trials: Estimands, Core Outcomes, and Adaptive Pandemic-Ready Frameworks	Frameworks for Clinical Research	Longitudinal and Joint Models of Clinical Trajectories	WS: From Research Question to Access: Using Nationwide Claims Data Today – and Electronic Health Record Data Tomorrow	Design and Analysis Choices in Clinical Trials	Missing Data and Measurement Error	Digital Health Tools and Engagement	WS: Developing and Executing Privacy-Preserving Federated Analyses with FLAME	Semantic Health Data and Meta-Research	WS: From Theory to Practice in the EU HTA JCA: Early Insights and Lessons Learned	
12:30 - 13:15			Empfang GMDS Neumitglieder	ISCB Student Conference Awards Committee	GMDS PK Ethische Fragen in der Medizinischen Informatik, Biometrie und Epidemiologie	GMDS AG Consumer Health Informatics (CHI)	GMDS PK Zertifikatskommission Biometrie in der Medizin	GMDS Fachausschuss Epidemiologie	GMDS AG Management von Informationssystemen im Krankenhaus (MIK)		GMDS AG Health Technology Assessment (HTA)	GMDS AG MI-Lehre in der Medizin	
12:30 - 13:30	Lunch Break												
13:30 - 15:00	IS: Causal Inference with Truncation by Death and Missing Data: From Estimands to Estimation	Survival Prediction Models	WS: Analysis and Modelling of Infectious Diseases in the Age of AI: Lessons from COVID-19	LLMs for Clinical Workflows	Learning Latent Clinical Trajectories	WS: Between Data, Practice, and Evidence: (Inter)Nationally EHR Research	Adaptive and Comparative Trial Design Strategies	Prediction and Imputation with Missing Data	Health Data Infrastructure and Secondary Data Use	WS:Mapping metadata schemas to the HealthDCAT-AP standard: Lessons learned and hands on community examples	Diagnostic Evidence Synthesis and Test Evaluation	WS: A Guided Experiment on Clinical Trial Data Sharing Quality and Data Reuse	
15:00 - 15:30	Coffee Break												
15:30 - 17:00	IS: Innovations in Master Protocols Evaluating Multiple Treatments and (Rare) Diseases	Dynamic Survival Modeling and Prediction	WS: Uncertainty in Clinical Predictions	Clinical Imaging	GMDS awards	WS: Large Language Model Informed Real-World Cohort Data	Adaptive Trial Inference and Error Control	Infectious Disease Modelling	Digital Health Education and Transformation	WS: Mapping metadata schemas to the HealthDCAT-AP standard: Lessons learned and hands on community examples	Network Meta-Analysis and Evidence Reliability	WS: A Guided Experiment on Clinical Trial Data Sharing Quality and Data Reuse	
17:00 - 18:30		GMDS Mitglieder-versammlung									GMDS Präsidiums-kommission Zertifikatskommission Medizinische Dokumentation		
18:30 - 20:30	Welcome Reception												

ISCB GMDS 2026

Preliminary Program (Day view)

BIG DATA small data Your Data
Transitions in the age of biomedical AI

Tuesday, 29.09.2026													
	Rolf Böhme Saal	Runder Saal	K1	K2	K3	K4	K5	N1	N2	N3	N4	N5	P1-P7
09:00 - 09:45	Keynote: Marylyn Ritchie												Partner & Poster-Exhibition
09:45 - 10:30	Keynote: Annette Kopp-Schneider												
10:30 - 11:00	Coffee Break												
11:00 - 12:30	IS: AI/ML-Driven Causal Inference with Real-World Data Complexities	Strategies for Prediction Models	WS: Reliable Subgroup Identification and Analysis - Results of GMDS Biostatistics Competition 2026	LLMs for Clinical Text and Data	Clinical Trajectories and Events	WS: Introduction to Individual Participant Data Meta-Analysis	Small Sample and Sequential Trial Methods	Robust Causal and Statistical Methods for Complex Data	Data Foundations for Clinical AI		Health Technology Assessment	WS: On the Ethical and Technical Aspects for Deployment and Evaluation of AI-Driven Clinical Decision Support Systems	
12:30 - 13:15			GMDS AG Entscheidungs-unterstützung im Gesundheitswesen	ISCB Education Subcommittee	GMDS AG Informations-verarbeitung in der Pflege	GMDS AG Statistische Methodik in der klinischen Forschung	GMDS & IBS AG Lehre und Didaktik der Biometrie	GMDS Projektgruppe Diversität	GMDS AG Curricula der Medizinischen Informatik (CMI)	GMDS Fachausschuss Medizinische Biometrie	ISCB SC Statistics in Regulatory Affairs (SiRA)	GMDS AG Medizinische Terminologien und Klassifikationen (MTK)	
12:30 - 13:30	Lunch Break												
13:30 - 15:00	Signature Session: AI and Me	Building and Updating Prediction Models	Data Quality, Selection, and Measurement	Statistical Learning under Uncertainty	Tailoring Statistical Models	Strategies for Evidence Synthesis	Adaptive Monitoring and Borrowing in Trials	Target Trial Emulation	Research-Ready Data Integration and Software Workflows	Prediction and Treatment	Sponsor Workshop: SAS Data Maker + Friedrich-Wingert-Stiftung	Treatment Effect Heterogeneity and Clinical Interpretation	
15:15 - End	Conference Excursions / Social Afternoon												

ISCB GMDS 2026

Preliminary Program (Day view)

BIG DATA small data Your Data

Transitions in the age of biomedical AI

Wednesday, 30.09.2026													
	Rolf Böhme Saal	Runder Saal	K1	K2	K3	K4	K5	N1	N2	N3	N4	N5	P1-P7
09:00 - 10:30	IS: Accelerating Evaluation and Adoption of Complex Digital Outcome Measures in Trials	WS: Ordinal Outcomes: Friend or Foe?	WS: AI-driven Therapeutic Modelling in Biomedical Informatics	WS: Benchmarking AI scientists	WS: Forum Digital Medicine and Artificial Intelligence	WS: Regulatory Topics for Statisticians: Implementation of Risk-Based Quality Management from Biostatistical Perspective	Dose Finding and Adaptive Trial Design	Causal Inference with Survival and Competing Events	Interoperable Data and Digital Evaluation	Omics Discovery	Strategies in Clinical Trials	Recurrent Events Modelling	Partner & Poster-Exhibition
10:30 - 11:00	Coffee Break												
11:00 - 12:30	IS: Biostatistics for Personalized Medicine	WS: Use of Bayesian Methods for Clinical Trial Design and Analysis	WS: AI-driven Therapeutic Modelling in Biomedical Informatics	Clinical Prediction	WS: Forum Digital Medicine and Artificial Intelligence	WS: Population-Adjusted Indirect Comparisons in the Framework of Health Technology Assessment	Basket Trials and Adaptive Borrowing	Causal Estimands and Biases	Federated Health Data Systems	WS: AI supported Data Retrieval and Data Management in the Health Domain	Diagnostic Estimands and Modeling	Advances in Multi-State Modeling	
12:30 - 13:15			GMDS Projektgruppe Global Health Informatics (GHI)	GMDS Projekt EVA4MII		GMDS AG Therapeutische Forschung	GMDS Arbeitskreis Epidemiologie der Arbeitswelt	GMDS AG Epidemiologie der Herz-Kreislauf- und Stoffwechsel-Erkrankungen	GMDS AG Medizinische Bild- und Signalverarbeitung (MBSV)	GMDS AG Mobile Informationstechnologie in der Medizin und Telemedizin (MoCoMed)	GMDS AG Informations-systeme im Gesundheitswesen KIS	GMDS Fachausschuss Medizinische Bioinformatik und Systembiologie	
12:30 - 14:00		AGM ISCB (12:30 - 13:40)	Lunch Break		WS: Forum Digital Medicine and Artificial Intelligence (13:00 - 14:00)	Lunch Break							
14:00 - 15:30	IS: Evaluating Fairness in Clinical Prediction Models and AI: Current and Emerging Topics	WS: Use of Bayesian Methods for Clinical Trial Design and Analysis	Platform and Cluster Trial Designs	Omics in Modeling	WS: Forum Digital Medicine and Artificial Intelligence	WS: What Makes a Trial Complex? Evidence, Experience, and AI	WS: Teaching Statistical Literacy in Times of Political Changes	Causal Estimands and External Controls	Real-World Evidence, Registries, and Surveillance	WS: Deep Dive Into eHealth With the Project MINT Meets Medicine	Cancer Screening	Time-to-Event Methods and Meta-Analysis	
15:30 - 16:00	Coffee Break												
16:00 - 16:45	Keynote: Andre Dekker												
16:45 - 17:30	Closing Session												
19:00 - End	Conference Dinner												

ISCB GMDS 2026

Preliminary Program (Day view)

BIG DATA small data Your Data
Transitions in the age of biomedical AI

Thursday, 01.10.2026			
	Rolf Böhme Saal	Runder Saal	K5
09:00 - 10:30	SmallData Symposium	STRATOS Symposium	Early Career Researcher Day
10:30 - 11:00	Coffee Break		
11:00 - 12:30	SmallData Symposium	STRATOS Symposium	Early Career Researcher Day
12:30 - 13:30	Lunch Break		
13:30 - 15:00	SmallData Symposium	STRATOS Symposium	Early Career Researcher Day
15:00 - 15:30	Coffee Break		
15:30 - 17:00		STRATOS Symposium	

Pre-Conference Courses

Sun 27.09.2026

Performing Comparative Effectiveness Research Using Network Meta-Analysis: An Introduction to Theoretical and Practical Advances

Sun, 27.09.26, 08:30–17:30, Room: K1

Course Instructor(s)

- **Theodoros Evrenoglou**

  Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center-University of Freiburg, Germany

- **Anna Chaimani**

  University of Oslo, Oslo Center for Biostatistics and Epidemiology, Department of Biostatistics, Norway

- **Guido Schwarzer**

  Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center-University of Freiburg, Germany

General outline and main topics

Network meta-analysis (NMA) is a comparative effectiveness research (CER) statistical method that enables the synthesis of evidence from multiple studies and the simultaneous comparison of multiple treatments for a given condition. By combining direct and indirect evidence, NMA can inform all possible treatment comparisons, including those for which head-to-head studies are unavailable. To date, NMA has become a standard tool in CER and is widely adopted by guideline developers, stakeholders, and major organizations such as the World Health Organization and the National Institute for Health and Care Excellence. Well-documented benefits of NMA include increased precision and reliability of treatment effect estimates, as well as the ability to translate evidence across multiple treatments into a single, coherent ranking. Recent methodological advances have extended NMA to more complex settings, including multiple-component interventions, multiple outcomes, and rare events which commonly arise in the analysis of safety outcomes.

Learning objectives

The aim of this course is to introduce both basic and advanced topics in NMA from methodological and applied perspectives. The theoretical part of the course will cover the foundations of several NMA approaches and illustrate their use and interpretation through real-world applications across diverse medical fields. In parallel, a series of interactive practical sessions will provide participants with hands-on experience in applying NMA methods and familiarity with the R packages meta, netmeta, and mvnma.

Structure

In the first part of the course, we will introduce the fundamentals of frequentist NMA, including model parameterization, estimation of relative treatment effects by combining direct and indirect sources of evidence, and treatment ranking. We will also explore the core NMA assumptions of transitivity and consistency, and their implications for the validity of results. In the second part, we will cover more advanced topics, including recent advances in treatment ranking extensions of NMA model to account for multiple-component interventions synthesis of evidence across multiple outcomes and the conduct of NMA in the presence of rare events. Finally, we will introduce NMAstudio and an AI-supported tool that enables translation of NMA findings into practice while leveraging large language models.

All lectures will be interactive and followed by hands-on practical sessions.

Appeal to ISCB GMDS 2026 attendees

By the end of the course, participants will have developed a solid understanding of the theoretical foundations of NMA, together with familiarity with recent methodological developments and tools. Through a combination of theory and hands-on practice, participants will acquire skills that enable them to critically assess, implement, and interpret NMA in applied settings. These competencies will allow participants to contribute effectively as statistical experts to clinical and public health applications, supporting evidence-based decision-making and health technology assessment.

Attendees expertise and technical requirements

This course is designed for an interdisciplinary audience, focusing on students and early career researchers with an understanding of biostatistics, as well as health professionals, policymakers, and epidemiologists interested in integrating evidence synthesis methods into decision-making. To fully engage with the practical sessions, participants are expected to have some experience with R.

Beyond "Randomized": Designing and Evaluating Randomization in Clinical Trials

Sun, 27.09.26, 08:30–12:30, Room: K2

Course Instructor(s)

- **Stefanie Schoenen**

  RWTH Aachen, Institute of Medical Statistics, Germany

- **Daniel Bodden**

  RWTH Aachen, Institute of Medical Statistics, Germany

- **Ralf-Dieter Hilgers**

  Sigmund Freud Private University Vienna, Austria

- **Franz König**

  Medical University of Vienna, Austria

General outline and main topics

Randomized controlled trials are widely regarded as the gold standard in clinical research because randomization and allocation concealment together protect against multiple sources of bias. Randomization aims to prevent systematic differences between treatment groups, while concealment protects assignments from foreknowledge and manipulation during enrolment.

In practice, however, the label "randomized" is often treated as a guarantee of validity, while the implementation is often overlooked. This is particularly problematic in open-label and increasingly complex trial designs, where randomization interacts with design elements such as interim analyses.

This course provides a comprehensive overview of randomization in clinical trials. It begins with a historical introduction and clarifies what randomization can (and cannot) achieve. Participants are introduced to different types of randomization procedures, including their strengths, limitations and the fundamental trade-off between predictability and imbalance.

A practical session demonstrates a Shiny app for evaluating and visualizing the impact of bias on inference, along with the R package randomizeR and the Evaluation of Randomization procedures for Design Optimization (ERDO) framework for simulation-based planning and generation of randomization lists. The course concludes by emphasizing that labeling a study "randomized" does not guarantee valid results and provides guidance on selecting appropriate randomization procedures and addressing challenges in complex trial designs.

Learning objectives

Participants will learn about the history and rationale of randomization in clinical trials, different randomization procedures with their strengths and limitations, and how to generate randomization lists. The course also covers how randomization affects evidence reliability, common pitfalls in randomization, and practical criteria for selecting an appropriate randomization procedure for a given study design.

Structure

Background

- Introduction to randomization in clinical trials
- Overview of randomization procedures and their trade-offs between predictability and imbalance
- Importance of transparent reporting of randomization
- Randomization-related issues affecting trial validity
- Selection criteria for randomization procedures in clinical trials, including complex designs such as adaptive, multicenter, and multiple-endpoint trials

Practical session

Hands-on demonstration of tools to analyze and visualize the impact of bias on inference for different randomization procedures and a framework for selecting appropriate randomization procedures for clinical trials.

- Demonstration of an R Shiny app for generating randomization lists and evaluating randomization procedures
- Presentation of key functionalities of the R package randomizeR for randomization list generation and evaluation according to the ERDO framework

Conclusion

- Summary of why "randomized" alone does not ensure validity
- Discussion of remaining challenges and issues in randomization, particularly for complex trial designs

Appeal to ISCB GMDS 2026 attendees

Randomized - but truly unbiased? Simply labeling a trial "randomized" does not guarantee valid results. As clinical trial design becomes more complex, the choice, implementation, and reporting of randomization becomes crucial. This session introduces the fundamentals of randomization, including a historical perspective, reviews different randomization procedures and discusses how these choices affect predictability, balance and vulnerability to bias and with that ultimately the level of evidence. Combining methodological insight with practical R-based tools (including a Shiny app), the session shows how to select randomization approaches that support more robust clinical trials.

Attendees expertise and technical requirements

Basic knowledge of randomized clinical trials and access to a digital device.

Efficient Bayesian Analysis of Longitudinal and Survival Data using R-INLA

Sun, 27.09.26, 08:30–12:30, Room: K4

Course Instructor(s)

- **Denis Rustand**

  University of Bordeaux, National Institute of Health and Medical Research, France

General outline and main topics

Modern biostatistical research is increasingly characterized by the analysis of complex, high-dimensional data from sources like electronic health records and large observational studies. While joint models provide a powerful framework for these data, their application have been limited by the prohibitive computational cost of traditional methods like MCMC. This course introduces a powerful computational framework: Integrated Nested Laplace Approximations (INLA). As a highly efficient alternative for Bayesian inference, INLA provides results of comparable or superior accuracy to MCMC but reduces computation time from hours to seconds or minutes. The content is structured around the comprehensive framework of the book 'Bayesian survival, longitudinal and joint models with INLA' (Chapman & Hall/CRC, May 2026). It showcases the framework's flexibility, enabling participants to implement multivariate joint models by combining various longitudinal distributions and handling complex survival scenarios like competing risks and multi-state models.

Learning objectives

Upon completion of this course, participants will be able to understand the fundamental principles of the INLA methodology and its advantages for biostatistical modeling. They will learn to formulate and fit a wide range of advanced Bayesian models for complex longitudinal, survival, and joint data structures. Participants will also be able to implement these models efficiently using the user-friendly INLAjoint R package. Finally, they will learn to generate and interpret dynamic predictions for personalized risk assessment, applying the fitted models for practical inference.

Structure

The course is structured in three parts. Part 1 (Foundational Concepts) introduces the motivation for joint modeling and provides a high-level overview of the INLA methodology and the INLAjoint package. Part 2 (Modeling with INLA) is dedicated to practical implementation, where participants will fit univariate survival and longitudinal models before combining them into multivariate joint models with various association structures. Part 3 (Prediction, Interpretation, and Wrap-Up) focuses on generating and visualizing dynamic predictions for longitudinal trajectories and survival probabilities, concluding with a Q&A and overview of advanced topics (e.g., non-linear associations, spatial autocorrelation).

Appeal to ISCB GMDS 2026 attendees

This course is highly relevant for statisticians analyzing complex observational studies or modern clinical trials with longitudinal biomarkers. It addresses the growing need to answer complex scientific questions where traditional methods may fail due to computational costs. By introducing a method that scales to large data and reduces computation time from hours to minutes, it makes joint modeling accessible for challenging datasets like registries. The free, open-source nature of the software and forthcoming book makes it accessible to researchers from any region.

Attendees expertise and technical requirements

This is an intermediate level course. Participants are expected to have a background in applied statistics and R programming. Specifically, they should possess a solid understanding of mixed-effects models, prior experience with survival analysis (proportional hazards), and basic experience using R. No prior experience with Bayesian statistics or INLA is required. Participants must bring a personal laptop with R installed to participate in the hands-on exercises.

Causal inference for time-to-event outcomes - with practical applications in R

Sun, 27.09.26, 08:30–17:30, Room: K5

Course Instructor(s)

- **Ruth Keogh**

  London School of Hygiene & Tropical Medicine, Medical Statistics Department, United Kingdom

- **Jon Michael Gran**

  University of Oslo, Oslo Centre for Biostatistics and Epidemiology, Norway

General outline and main topics

Recent years have seen major developments in methods for causal inference using observational data. However, the practical application of the methods is challenging and lags behind methodological developments. This is especially true in the context of survival and other time-to-event outcomes, which are commonly of interest in applications in biostatistics and data science.

This course will provide training on concepts and methods for estimating causal effects of treatments on time-to-event outcomes. It will begin with an introduction to causal estimands and assumptions required for their identification using observational data. It will then cover estimation methods for confounding adjustment, including inverse probability weighting, marginal structural models, g-formula, and censoring-weighting approaches. The initial focus will be on treatments given at a single time point, before extending to time-varying treatment strategies. Methods that incorporate machine learning techniques will be included, and extensions to settings with competing events will also be discussed.

The course material will be presented with medical and epidemiological applications in mind, but the methods are equally relevant in other areas, such as social science and economics. The course will combine lectures and computer practical sessions using openly accessible data sets.

Learning objectives

Select causal estimands suitable for time-to-event outcomes, including when there are competing events. Apply methods for estimating the effects of point treatments and longitudinal treatment strategies in R, including inverse probability weighting, g-formula and doubly-robust methods using machine learning techniques. Understand the assumptions, advantages and disadvantages of different estimation methods.

Structure

The course will comprise interactive lectures and computer practicals. Practicals will be guided by exercise sheets and include discussion of solutions.

Lecture 1: Causal estimands and identification using observational data Lecture 2: Estimating the effects of point treatment strategies Computer practical 1 Lecture 3: Methods for time-dependent

treatment strategies Computer practical 2 Lecture 4: Doubly robust methods Computer practical 3 Lecture 5: Extensions and summary

Appeal to ISCB GMDS 2026 attendees

This course is aimed at researchers and students in biostatistics, epidemiology and data science. Causal inference methods are increasingly required in these fields, with recent developments emphasising use of machine learning techniques. Time-to-event outcomes are very commonly faced but training focused on this aspect has previously been lacking. The course will be of interest to participants wishing to gain a comprehensive overview of methods in this area and to those focused on applications.

Attendees expertise and technical requirements

Some knowledge of methods for survival analysis will be assumed, such as estimation of survival curves using the Kaplan-Meier estimator and fitting regression models. Knowledge of basic causal concepts such as confounding would be beneficial. Participants should have prior experience of using R, though not necessarily for time-to-event analysis or causal inference. In advance of the course participants will be provided with background reading on some fundamental methods for time-to-event and on key causal concepts. We will make available a brief online tutorial enabling participants to familiarise themselves with the main functions for survival analysis in R.

Prediction Under Interventions

Sun, 27.09.26, 13:30–17:30, Room: K2

Course Instructor(s)

- **Nan van Geloven**

  Leiden University Medical Center, The Netherlands

- **Karla Diaz-Ordaz**

  University College London, United Kingdom

- **Doranne Thomassen**

  Leiden University Medical Center, The Netherlands

General outline and main topics

A key aim of prediction algorithms is to provide individualized risk estimates that support end users in making decisions on interventions, for example personalizing a medical treatment. Most prediction models are however derived from observational training data where some individuals already receive the interventions the model aims to inform. This makes standard prediction models unfit to provide actionable information.

In this half-day course we focus on methods for developing and evaluating predictions under interventions. These are estimates of risk under the specific intervention options that users need to decide on and can be obtained by combining prediction methods with causal inference techniques.

Learning objectives

Participants will learn to

- identify pitfalls of using regular predictions for decision support

- implement methods for developing interventional prediction algorithms

- implement methods for evaluating counterfactual performance of predictions

Structure

- Causal blind spots in risk-based decision making (30 min lecture)

- Train a regular prediction algorithm to a case study dataset + understand its shortcomings (30 min exercise in R)

- Developing interventional prediction algorithms, explained using of variety of causally-aware methods, such as g-computation, (a)ipw and debiased machine learning (45 min lecture)

- Train interventional prediction algorithm on the case study data (40 min exercise in R)

BREAK (20 min)

- Evaluating counterfactual performance of a prediction model in absence of ground truth, i.e. in observational data (45 min lecture)

- Evaluate the regular and the interventional prediction algorithm against a counterfactual data distribution (30 min exercise in R)

Appeal to ISCB GMDS 2026 attendees

Bridging the gap between predictive modeling and personalized medicine is a shared challenge for bioinformaticians, biostatisticians, and epidemiologists. This course explores the essential intersection of machine learning and causal inference. As such, it provides an excellent kick-off for further cross-disciplinary exchange among the diverse communities attending ISCB GMDS 2026.

Attendees expertise and technical requirements

We assume participants are versed in regular prediction techniques: training a regression or machine learning model, and assessing its performance with model fit measures such as RMSE or AUC. We also assume familiarity with using R (though code for carrying out the practical exercises will be provided). We assume only basic familiarity with causal inference concepts such as confounding, other necessary concepts will be introduced.

Clinical Text Mining under Real-World Constraints

Sun, 27.09.26, 13:30–17:30, Room: K3

Course Instructor(s)

- **Vittorio Torri**

  MOX - Modelling and Scientific Computing Lab, Department of Mathematics, Politecnico di Milano, Italy

- **Francesca Ieva**

  MOX - Modelling and Scientific Computing Lab, Department of Mathematics, Politecnico di Milano, Italy; HDS - Health Data Science Centre, Human Technopole, Italy

General outline and main topics

Statistical and machine learning models for healthcare analytics increasingly rely on extracting structured variables from unstructured clinical documents such as discharge summaries and clinical notes. While Natural Language Processing (NLP) has advanced rapidly in recent years, its application to real-world medical data faces relevant domain-specific challenges. The limited availability of gold-standard annotated data, due to high annotation costs, and the constrained computational resources in hospital environments restrict the applicability of many state-of-the-art NLP models. Moreover, privacy regulations can complicate data sharing and large-scale model development. These constraints create a growing need for methodologies enabling reliable information extraction under realistic resource and data constraints.

This course provides a practical overview of such techniques. The course will cover robust design and validation of rule-based approaches, unsupervised representation learning and clustering, weakly supervised classification techniques, and data augmentation methods based on Large Language Models (LLMs). For each topic, we will introduce the methodological foundations, present a real-world case study (CS) where we applied the approach, and provide a hands-on programming exercise (EX). Exercises will use publicly available clinical text resources in English, including documents from the European Clinical Case Corpus (E3C) and from a dataset released for the CRF Filling Shared Task at CL4Health - LREC 2026 (FBK-CRF).

Learning objectives

By the end of the course, participants will understand the main methodological options for extracting structured information from clinical text when labelled data is limited. They will learn how to select and implement workflows ranging from rule-based systems to large language models, based on the problem, the availability of labelled and unlabelled data, and computational resources. They will also learn how to evaluate these methods using small but representative labelled datasets, including the definition of annotation guidelines and the measurement of inter-annotator agreement.

Structure

Introduction (15 minutes) Rule-based methods - CS: Assessing waiting times for follow-up exams from referrals - EX: Extracting lab measurements (E3C) (45 minutes) Clustering - CS: Evaluating referral appropriateness - EX: Diagnoses clustering (E3C) (45 minutes) Weak supervision - CS: Phenotyping bronchiolitis and heart failure from discharge summaries - EX: Identifying dyspnoea-related notes (FBK-CRF) (1 hour) Break (15 minutes) Data Augmentation with LLMs - CS: Structuring EHR fields for STEMI patients - EX: Diagnoses classification (E3C) (45 minutes) Q&A (15 minutes)

Appeal to ISCB GMDS 2026 attendees

The course directly addresses the growing need to integrate unstructured clinical data into statistical and epidemiological analyses. For ISCB attendees, it provides practical tools for cohort identification, phenotype classification, and extraction of structured variables usable in standard regression and survival models. For GMDS attendees, it is particularly relevant because it addresses clinical data integration challenges and practical deployment constraints that are common in medical informatics.

Attendees expertise and technical requirements

Basic knowledge of programming (preferably Python), statistics, and machine learning. Participants should bring a laptop with internet access and a Google account to run Colab Notebooks.

Genomic AI Agents for Personal Genomics: Hands-on Analysis with Just-DNA-Lite and LLM agents

Sun, 27.09.26, 13:30–17:30, Room: K4

Course Instructor(s)

- **Anton Kulaga**

  IBIMA, Institute for Biostatistics and Informatics in Medicine and Ageing Research University Medical Centre Rostock, Germany; Institute of Biochemistry of the Romanian Academy (IBAR), Romania

- **Livia Zaharia**

  HEALES (Healthy Life Extension Society), Belgium

- **Nikolay Usanov**

  HEALES (Healthy Life Extension Society), Belgium

General outline and main topics

This course teaches participants to analyze genomes using open-source tools and AI agents. We cover the pipeline from FASTQ through alignment and variant calling to VCF in theory, then focus hands-on on annotation and reporting, since alignment takes hours. Participants run Just-DNA-Lite locally to produce disease-risk and polygenic risk score (PRS) reports, filter annotations, then use AI agents to build a new annotation module from scientific literature and apply it to real genomic data. Longevity genomics serves as the running example. Time permitting, we demonstrate variant effect prediction with genomic foundation models.

Learning objectives

Participants will: (1) understand the genomic pipeline from FASTQ through VCF to annotated reports (theory only) (2) run Just-DNA-Lite locally to produce disease-risk and PRS reports (3) filter and interpret annotation outputs (4) use AI agents to create a new annotation module from published literature and polygenic risk scores (5) apply that module to genomic data and critically evaluate the results.

Structure

Part 1 - Genomics Background (20 min): Pipeline overview (FASTQ → BWA-MEM2 → DeepVariant → VCF → annotation, theory only), variants and disease risk, GWAS/PRS, longevity case study (APOE, FOXO3, CETP existing PRS and heritability estimates).

Part 2 - Running Just-DNA-Lite & Exploring Annotations (60 min): Install the platform, upload VCFs (instructors' genomes, NA12878, or own), run longevity/PRS/health-risk pipelines, explore Dagster UI and generated reports, filter Parquet outputs with Polars/DuckDB, cross-reference ClinVar, gnomAD, functional scores.

Break (20 min).

Part 3 - Building a Module with AI Agents (80 min): Participants supply papers or a PRS from PGS Catalog the agent generates an annotation module.

Part 4 - Applying Your Module (30 min): Run AI-built module on real data, compare results, discuss curation.

Part 5 (if time) - Foundation Models (25 min): Advanced regulatory variant effect prediction (AlphaGenome, NucleotideTransformer etc.)

Appeal to ISCB GMDS 2026 attendees

The conference theme 'Big Data, Small Data, Your Data' fits exactly: participants analyze their own genomic data using large-scale databases and AI. The course bridges biostatistics (PRS methodology), medical informatics (clinical variants), and bioinformatics (annotation pipelines). Participants do not just run pre-built analyses - they use AI agents to create a new module from literature and test it on real genomes. All software is open-source.

Attendees expertise and technical requirements

Researchers and students with basic biology knowledge and IT skills. Life-science background is not required - IT expertise alone suffices. No AI/LLM experience needed. Participants bring a laptop (Linux, macOS, or Windows with WSL) with Python 3.12+ and uv pre-installed. Those with sequenced genomes can bring their VCF file.

Opening Session - Welcome to ISCB GMDS 2026

Join us as we officially open the conference and welcome our community to Freiburg!

Welcome Messages

The session will begin with welcome and opening remarks from **Harald Binder**, Chair of the Local Organizing Committee (LOC), and **Nadine Binder**, Chair of the Scientific Program Committee (SPC). They will be followed by addresses from **Lutz Hein**, Dean of the Faculty of Medicine, University of Freiburg, and **Frederik Wenz**, Chief Medical Officer of the Medical Center – University of Freiburg.

ISCB and GMDS Awards Celebration

The Opening Session will also recognize the achievements of members of our scientific community. **Andrea Jorgensen**, Chair of the ISCB Student Conference Awards Subcommittee, will present the ISCB Awards, followed by **Hans-Ulrich Prokosch**, President of GMDS, who will present the GMDS Awards. The award presentations are an opportunity to acknowledge the work and contributions of colleagues and young researchers within our community.

Keynote Sessions

Cathie Sudlow: Future of Population Health Research: a UK Health Data Perspective

Mon, 28.09.26, 09:45–10:30, Room: Rolf Böhme Saal

Chair: Nadine Binder

Keynote Profile

Cathie Sudlow
Usher Institute and School for Population Health Sciences, University of Edinburgh, Scotland, UK

Cathie Sudlow is Professor of Neurology and Clinical Epidemiology, Director of the Usher Institute, Head of School for Population Health Sciences at the University of Edinburgh and Director of the UKRI Adolescent Health Study (AHS). Moreover, she is a Fellow of the Academy of Medical Sciences, the Royal Society of Edinburgh and Honorary Fellow of Health Data Research UK. She was awarded an OBE for services to medical research in 2020. Her research interests have always been firmly embedded in the world of big data. Over the last 15 years, her focus has been on leading large-scale, collaborative, open-science initiatives that enable a better understanding of the causes and consequences of health and disease across the life course, leading to new and improved approaches to prevention, diagnosis and treatment. From 2011 to 2019, as Chief Scientist for UK Biobank, she led efforts to follow the health of UK Biobank participants through linkage to national health datasets. From 2020, she worked with NHS Digital and then NHS England to develop NHS England's first secure data environment to hold and enable access for research to linked health data from multiple sources for the whole population of England. In 2023, Sudlow was commissioned by the Chief Medical Officer for England, the UK National Statistician and NHS England to undertake an independent review of the UK-wide health data landscape.

Marylyn Ritchie: The End of Data-Driven Medicine: Entering the Era of AI-Native Science

Tue, 29.09.26, 09:00–09:45, Room: Rolf Böhme Saal

Chair: Nadine Binder

Presenter/Author: Marylyn Ritchie

Affiliations: Medical University of South Carolina, Charleston

For decades, biomedical research has been built on a familiar model: collect data, formulate hypotheses, analyze results, and translate discoveries into practice. Artificial intelligence is not simply accelerating this process-it is fundamentally changing it. We are entering an era in which AI systems can integrate multimodal evidence across genomes, health records, images, sensors, literature, and scientific knowledge at a scale that exceeds human cognitive capacity. As a result, the central challenge facing biomedical science is no longer whether we can generate enough data. The challenge is whether our institutions, methods, and governance structures are prepared for a future in which discovery itself becomes increasingly machine-assisted. This talk will argue that the age of biomedical AI represents a transition from data-driven medicine to AI-native science. In this new paradigm, intelligent systems will not merely identify patterns in data but will actively participate in generating hypotheses, designing experiments, predicting outcomes, identifying intervention opportunities, and supporting clinical decision-making. The traditional boundaries between research, clinical care, and operational improvement will begin to dissolve as learning health systems become capable of continuously converting experience into evidence and evidence into action. Yet the greatest barriers to this future are not technical. They are human. Healthcare organizations remain structured around workflows, incentives, governance models, and educational frameworks that were designed for a pre-AI world. Success will depend on our ability to develop trustworthy systems that augment-not replace-human expertise while preserving transparency, accountability, scientific rigor, and equity. Drawing on perspectives from precision medicine, biomedical informatics, and enterprise AI leadership, this presentation will explore what must change if we are to realize the promise of AI-enabled discovery and healthcare transformation. It will challenge attendees to rethink the future role of scientists, clinicians, and health systems in a world where machines increasingly contribute to the generation of knowledge itself. The most important question facing our field is no longer Can AI analyze biomedical data? It is How will biomedical science evolve when intelligence becomes an integral component of the scientific process?

Keynote Profile

Marylyn Ritchie

Biomedical Informatics and AI, Medical University of South Carolina, College of Medicine, Charleston, SC, USA

Marylyn D. Ritchie is Chief Artificial Intelligence Officer for the MUSC Enterprise, Director of the Center for AI, Associate Dean for Artificial Intelligence and Director of the Division of Computational Health Sciences and AI in the College of Medicine at the Medical University of South Carolina. She is also the SmartState Endowed Chair in Translational Biomedical Informatics. Dr. Ritchie is an expert in translational bioinformatics, with a focus on developing, applying and disseminating algorithms, methods and tools integrating electronic health records with genomics. She has over 20 years of experience in translational bioinformatics and has authored over 500 publications. Dr. Ritchie was appointed as a Fellow of the American College of Medical Informatics in 2020 and elected as a member of the National Academy of Medicine in 2021. Moreover, Dr. Ritchie is dedicated to promoting the implementation of precision medicine into routine clinical care through research programs focused on pharmacogenomics, genomic medicine and the learning health system.

Annette Kopp-Schneider: Leveraging External Data in Clinical Trials: Principles, Perils & Potential

Tue, 29.09.26, 09:45–10:30, Room: Rolf Böhme Saal

Chair: Thomas Jaki

Presenter/Author: Annette Kopp-Schneider

Affiliations: German Cancer Research Center (DKFZ), Heidelberg

In the era of precision medicine, clinical trials are often conducted in small patient cohorts, investigating targeted therapies in molecularly characterized patients. These trials, however, do not occur in isolation, and substantial experience exists, for example, with standard-of-care treatments to which targeted therapies are compared. Consequently, external information borrowing can be considered to improve a trial's efficiency. The Bayesian paradigm is ideally suited for this aim, since external information can be incorporated into an informative prior for the current analysis. A potential issue with this procedure is that external and current information may conflict, and such inconsistency may not be predictable a priori. To protect against the harmful impact of potentially conflicting external information, a widely adopted robustification strategy is to dynamically discount the amount of transferred information based on the discrepancy between external and current data. The recently published FDA draft guidance on the use of Bayesian methodology in clinical trials marks a major step toward broader acceptance of Bayesian methods in regulatory submissions. Nevertheless, the operating characteristics of trial designs are crucial for understanding how trial data support correct conclusions. In this talk, frequentist and Bayesian operating characteristics of trial designs with borrowing from external data will be discussed. It will be shown that, in general, no power gains are possible if strict control of the frequentist type I error rate is desired, and such a result also applies to any robust dynamic method. However, the frequentist approach may be unrealistically stringent, and the Bayesian framework may be more appropriate, particularly in precision medicine, since benefits can be obtained if external information is reliable and consistent. This talk will provide an overview of methods for borrowing from external information, discuss metrics to assess their performance and illustrate the potentials and limitations of using external data. Co-authors: Manuel Wiesenfarth, Vivienn Weru, Silvia Calderazzo

---

Keynote Profile

Annette Kopp-Schneider

Biostatistics, German Cancer Research Center (DKFZ), Heidelberg, Germany

Annette Kopp-Schneider is head of the division of Biostatistics at the German Cancer Research Center. She holds a doctorate in mathematics and computer science from RWTH Aachen. Her research interests range from stochastic modeling of cellular processes and dose-response modeling to clinical statistics. Her involvement as biostatistician for the pediatric oncology trial series INFORM2 has started her interest in precision oncology and Bayesian clinical trial design. Her current focus is on the use of external information in clinical trials and the impact on trial operating characteristics. She served as president of the German Region of the International Biometric Society from 2021 to 2022.

Andre Dekker: AI for Better Health Care

Wed, 30.09.26, 16:00–16:45, Room: Rolf Böhme Saal

Chair: Nadine Binder

Presenter/Author: Andre Dekker

Artificial intelligence (AI), big data, and data science are rapidly reshaping health care, offering new opportunities to improve outcomes, increase efficiency, and enable more personalized medicine. This contribution will explore how AI can move beyond technological promise to deliver real impact in clinical practice, with a particular focus on oncology as a leading domain of innovation. AI has the potential to transform health care along two major dimensions. First, it can enhance efficiency by automating routine tasks, streamlining workflows, and reducing administrative burden. Second, and more importantly, AI can improve effectiveness by supporting clinical decision-making, predicting outcomes, and tailoring treatments to individual patients. These capabilities are central to the transition toward precision medicine. A key prerequisite for successful AI is access to large, sufficient-quality and diverse datasets. However, health care data are often fragmented across institutions and constrained by privacy, ethical, and regulatory considerations. As a result, much of the available data remains underused. In response, new paradigms such as FAIR data principles and federated learning infrastructures have emerged. These approaches enable analysis across distributed datasets while preserving patient privacy, allowing institutions to collaborate without sharing sensitive data. Using examples such as the Personal Health Train and other federated data networks, the contribution will demonstrate how privacy-preserving, multi-center collaboration can accelerate knowledge generation and support continuous learning from routine clinical data. This "rapid learning" approach enables health systems to evolve based on real-world evidence, bridging the gap between research and practice. Despite these advances, significant challenges remain. Translating AI models into clinical benefit requires robust validation, integration into existing workflows, and acceptance by clinicians and patients. Without careful implementation, AI risks remaining a promising technology without meaningful impact. Ultimately, the goal of AI in health care is not simply to generate more data or more complex models, but to improve patient outcomes and quality of life. Achieving this requires a combination of technological innovation, responsible data governance, and global collaboration. This contribution will provide both a conceptual framework and practical examples to guide the implementation of AI for better health care.

Keynote Profile

Andre Dekker
Maastro Clinic, Maastricht University Medical Centre+ and Maastricht University, The Netherlands

Andre Dekker is a Full Professor of Clinical Data Science and a board-certified Medical Physicist and is currently the Head of Clinical Data Science at Maastricht University, Maastricht UMC+ and Maastro Clinic. He also holds the position of Chief Scientific Officer at Medical Data Works. Dekker's research centers on federated FAIR data infrastructures, AI for health outcome prediction models, and applying AI to enhance patient and citizen health. He has published over 150 peer-reviewed articles and patents and is co-author of the book Fundamentals of Clinical Data Science.

Invited Sessions

Mon 28.09.2026

11:00–12:30

Biomedical Knowledge Integration

Mon, 28.09.26, 11:00–12:30, Room: Rolf Böhme Saal

Chair: Dagmar Waltemath, Thomas Ganslandt

Organizers: Dagmar Waltemath, Thomas Ganslandt

The increasing amount of data being generated in healthcare demands new technologies for data organisation, integration, management and exploration. Semantic Web Technologies and Knowledge Graphs have proven to be suitable approaches to support health data management, and the definition of FAIR metadata improves reusability of health data records. In this session we will discuss latest developments and applications in the field.

Speakers

- **Michel Dumontier**: Maastricht University, Institute of Data Science, Department of Advanced Computing Sciences, Netherlands
- **Sabine Oesterle**: SIB Swiss Institute of Bioinformatics, Switzerland
- **Maria-Esther Vidal**: Leibniz University Hannover, Leibniz Information Centre for Science and Technology, Germany

---

The AI in FAIR: New Approaches to Biomedical Data Interoperability

11:00-11:30

Presenter/Author: Michel Dumontier

Affiliations: Maastricht University, Maastricht

Biomedical research increasingly depends on combining data from multiple sources, including clinical studies, biobanks, and public repositories. However, differences in formats, structure, and terminology often make data difficult to integrate and reuse. The FAIR data principles - Findable, Accessible, Interoperable, and Reusable - provide high-level guidance for improving data management, but translating these principles into practice remains a challenge. This talk presents a practical approach to FAIR, focusing on how researchers can make their data easier to understand, validate, and combine, supported by AI-assisted tooling. Central to this is the use of shared vocabularies and simple data models that provide a common structure across datasets. I will illustrate how lightweight semantic frameworks such as the Semanticscience Integrated Ontology and the Simplified Upper Level Ontology help different datasets "speak the same language," and how constraint languages like ShEx and SHACL support data quality by checking that data conforms to expected structure and intended meaning. These ideas are demonstrated through two case studies: a large-scale knowledge base integrating metagenomics, geochemistry, and environmental data, and a healthcare use case focused on enabling interoperability across clinical data standards. Finally, I highlight the growing role of artificial intelligence in supporting FAIR data practices. AI-assisted tools can help automate data and concept mapping, suggest appropriate terminology, and scale the integration of heterogeneous datasets. Together, FAIR principles, semantic modeling, and AI enable more seamless and scalable biomedical data interoperability across studies and systems.

What Can We Really Do with Biomedical Knowledge Graphs? From FAIR Data to Explainable AI

11:30-12:00

Presenter/Author: Maria-Esther Vidal

Affiliations: TIB- Leibniz Institute for Science and Technology, Hannover; Leibniz Universität Hannover, Hannover

Biomedical knowledge graphs are widely regarded as a promising approach for integrating heterogeneous data. However, moving from FAIR data to AI systems that effectively support scientific research requires more than integration-it demands a structured process that makes data interpretable, traceable, and actionable. In this talk, I present how Semantic Web principles enable this transition. Starting from FAIR data management, metadata describing scientific artifacts-such as publications, datasets, projects, and computational models-is systematically captured and interlinked. Infrastructures such as the Leibniz Data Manager support the consistent representation and sharing of FAIR metadata across institutions, while platforms such as the Open Research Knowledge Graph (ORKG) enable the structured representation of scientific knowledge. Through ontology-driven modeling and declarative mappings, this metadata is transformed into semantically consistent knowledge graphs that form part of a broader knowledge graph ecosystem. Within this ecosystem, lifecycle management and constraint-based validation play a central role. Constraints ensure data quality and consistency, while lifecycle-aware processes support the continuous evolution, enrichment, and governance of knowledge graphs, resulting in reliable and interoperable representations of scientific knowledge. On top of this foundation, AI methods-particularly

neuro-symbolic approaches-leverage structured and validated knowledge to support scientific tasks such as literature-grounded hypothesis generation, link prediction across biomedical entities, and evidence-based reasoning. Because predictions are grounded in semantically integrated and validated metadata, the resulting AI systems are more transparent, explainable, and reproducible. This end-to-end perspective demonstrates how combining Semantic Web concepts, knowledge graph ecosystems, and AI methods enables a principled transition from FAIR data to explainable AI, ultimately supporting researchers and clinicians in navigating complex biomedical knowledge and accelerating discovery.

The SPHN Federated Clinical Routine Dataset as a Blueprint for Scalable Real-World Health Data Reuse

12:00-12:30

Presenter/Author: Sabine Oesterle

Affiliations: SIB Swiss Institute of Bioinformatics, Basel

The Swiss Personalized Health Network (SPHN) is a national data infrastructure funded by the Swiss Confederation to enable the systematic, FAIR-aligned reuse of health-related data for research and innovation, with the long-term objective of advancing personalized medicine in Switzerland. SPHN addresses key challenges in health data reuse-fragmentation, semantic heterogeneity, and limited interoperability-through a coordinated, federated framework spanning Swiss university hospitals and research institutions. Central to SPHN's approach is the secondary use of real-world data generated in routine clinical care and research, including electronic health records, laboratory data and medical imaging. The SPHN Federated Clinical Routine Dataset currently comprises data from more than 800,000 patients with broad consent across six university hospitals, enabling large-scale, multicenter research grounded in real-world clinical practice. Rather than centralizing sensitive patient-level data, SPHN implements a federated architecture in which data are locally curated, semantically annotated, and stored as institutional knowledge graphs. These graphs collectively contain more than 12.5 billion semantic triples mapped to 125 harmonized semantic concepts covering a broad spectrum of clinical and biomedical domains. This design embeds semantic interoperability at the source, transforming heterogeneous hospital data into an evolving, interoperable national research asset. FAIR principles are operationalized through harmonized data models based on international terminologies, machine-readable metadata, and a dedicated tool stack supporting both data providers and data users. Discoverability is enabled via a metadata-driven access model, where descriptive metadata and summary statistics are published in the SPHN Metadata Catalog, aligned with HealthDCAT and the European Health Data Space. Secure data access and analysis are performed through trusted research environments, ensuring compliance with legal, ethical, and governance requirements. Together, SPHN's federated architecture, semantic depth, and metadata-centric data discovery model demonstrate how FAIR principles can be translated into deployable infrastructure components at national scale, providing a practical blueprint for interoperable, health data ecosystems.

13:30–15:00

Causal inference with truncation by death and missing data: from estimands to estimation

Mon, 28.09.26, 13:30–15:00, Room: Rolf Böhme Saal

Chair: Cécile Proust-Lima, Margarita Moreno-Betancur

Organizers: Cecile Proust-Lima, Margarita Moreno Betancour, Stijn Vansteelandt

The estimation of causal effects is often hampered by the presence of incomplete data either due to death or other reasons. Incompleteness raises challenges across all steps of causal inference: estimand definition, identifiability, and estimation. This session aims to showcase methodological innovations to tackle incomplete data across the three key steps.

Firstly, the challenges of defining an estimand in the context of truncation by death will be discussed with a dual perspective (pragmatic evaluation in clinical research and mechanistic insights into disease processes). Solutions proposed in recent years include the survivor average causal effect, composite estimands such as survival-incorporated quantiles or the win ratio, and hypothetical estimands inspired by causal mediation analyses.

Secondly, recent developments regarding identifiability of causal effects in multivariable missingness settings depicted by so-called "missingness" directed acyclic graphs (DAGs) will be exposed. Finally, innovative estimation techniques will be addressed through causal machine learning methods for missing data.

Speakers

- **Jessica Young**: Harvard T.H. Chan School of Public Health, Department of Epidemiology, USA
- **Ghazaleh Dashti**: Murdoch Children's Research Institute & University of Melbourne, Australia
- **Matthew Pryce**: London School of Hygiene & Tropical Medicine, Department of Medical Statistics, UK

Story-led causal inference in truncation by death settings

13:30-14:00

Presenter/Author: Jessica Young

Affiliations: Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston

In a "truncation by death setting", an average causal effect (ACE) relative to the original study population, the default effect notion for causal inference, is undefined. In turn, the statistician guiding a subject matter investigator on data analysis decisions cannot default to any ACE as the foundation for those decisions another causal effect notion that is defined in the presence of deaths during the outcome period must be selected. Many alternative effect notions have been considered, including effects under "elimination of death" (which amounts to conceptualizing death as a censoring event), the survivor average causal effect or SACE (the ACE in the principal stratum of "always survivors"), and the "survival-incorporated quantile effect" (an effect on a ranked composite outcome). Interestingly, no matter what estimand is advocated, the literature on truncation by death and estimand choices takes for granted a setting where the assumption of treatment-death isolation (informally, "no treatment effect on survival") fails. When treatment-death isolation fails, a "naive" comparison of (counterfactual) average outcomes among survivors - what we will term a "survival conditioned contrast" (SCC) - can be identified by design in an ideal trial but cannot, in general, be interpreted as a causal effect (a comparison of outcomes under different tretament strategies across comparable populations). However, when treatment-death isolation holds, the SCC is a causal effect in this sense. In this talk, I will consider several possible underlying "causal stories" that typically motivate an investigator to specifically express interest in a treatment effect on an outcome that is truncated by death (instead of expressing interest in a treatment effect on survival). In all of these examples, I will argue that treatment-death isolation is not only a reasonable assumption, but a premise of the investigator's decision to focus on this outcome and not survival. Finally, I will highlight the difference between choosing the causal target estimand as the SCC versus analyzing a data set in a way that pre-conditions on those who survive the study period. I will also discuss the connection relation between the SCC and a conditional separable effect.

Handling multivariable missing data in causal inference

14:00-14:30

Presenter: S Ghazaleh Dashti

Authors: S Ghazaleh Dashti [1]; Katherine Lee [2]; John B. Carlin [1]; Margarita Moreno-Betancur [1]

Affiliations: [1] Murdoch Children's Research Institute and University of Melbourne, Melbourne; [2] Murdoch Children's Research Institute, Melbourne

Recent decades have seen major advances in conceptual frameworks and methods for causal inference, but there has been relatively little work on how to address the widespread problem of multivariable missing data in this context. Despite the central role of assumptions in formal causal inference, the additional assumptions required when data are missing have received scant attention. Further, there is insufficient guidance on appropriate approaches for handling missing data at the causal effect estimation stage, which are needed not only to obtain unbiased estimates but also to improve precision. This talk will provide an overview of the use of missingness graphs, or directed acyclic graphs (DAGs), to assess assumptions about multivariable missingness, as well as emerging work applying this approach in epidemiological settings to guide the handling of missing data. First, it

will introduce research that has developed a set of missingness DAGs depicting typical missingness mechanisms in real-world epidemiological studies, and that has derived theoretical results demonstrating implications for causal effect identifiability. Second, it will describe findings from extensive simulation studies investigating the performance of methods for handling missing data under these mechanisms, including examination of how to tailor multiple imputation procedures in the context of modern causal analytic methods. The talk will conclude by describing a structured framework developed to facilitate the translation of these ideas into epidemiological practice.

Causal machine learning for estimating heterogeneous treatment effects when outcome data is incomplete.

14:30-15:00

Presenter: Matthew Pryce

Authors: Matthew Pryce [1]; Stijn Vansteelandt [2]; Karla Diaz-Ordaz [1]; Ruth Keogh [3]

Affiliations: [1] University College London, London; [2] Ghent University, Ghent; [3] London School of Hygiene & Tropical Medicine, London

Machine learning offers researchers the opportunity to capture complex, high-dimensional relationships in data without having to pre-specify parametric forms. Yet machine learning algorithms are typically designed for predictive problems and do not naturally generate causal inferences. For this reason, much work has been done on developing causal frameworks that incorporate machine learning, allowing for the use of these tools to estimate causal estimands. In this talk, I discuss how causal machine learning estimators have been developed for exploring heterogeneous treatment effects (HTEs), with a particular focus on the challenges that arise when outcome data is incomplete. The main aim of this talk is to highlight how causal ML estimators can be extended to robustly estimate HTEs when the outcome of interest is censored or missing, focusing on two estimands suited to different data structures: the conditional average treatment effect (CATE), appropriate for binary or continuous outcomes, and the difference in conditional survival probabilities, which targets time-to-event outcomes. For the CATE, I examine how two prominent causal machine learning estimators, the DR-learner and the EP-learner, can be extended to handle missing-at-random outcome data. I then turn to the time-to-event setting, where incomplete outcome data arises inherently through right-censoring. In this setting, I introduce a newly developed estimator, surv-iTMLE, which offers a targeted learning based estimation approach for the difference in conditional survival probabilities. In both settings, I aim to highlight how these estimators robustly handle outcome missingness/censoring, and further, offer practical guidance on how to implement these methods in realistic clinical data settings.

15:30–17:00

Innovations in Master Protocols Evaluating Multiple Treatments and (Rare) Diseases

Mon, 28.09.26, 15:30–17:00, Room: Rolf Böhme Saal

Chair: Martin Posch, Stefan Michiels

Organizers: Stefan Michiels, Martin Posch

The traditional fragmented clinical development of treatments are being optimised by master protocols, which are set up to evaluate multiple treatments and (rare) diseases in the same trial. The session will cover recent statistical developments in early to late phase platform trials, precision medicine designs such as basket and umbrella trials used extensively in oncology, and multi-arm adaptive trials, with both Bayesian and frequentist approaches.

This session will be highly valuable for conference participants working on trials.

Speakers

- **Jack J Lee**: The University of Texas MD Anderson Cancer Center, Houston, TX, USA
- **James Wason**: Newcastle University, Population Health Sciences Institute, UK
- **Sofia Villar**: University of Cambridge, MRC Biostatistics Unit, UK

The successes and challenges of Bayesian adaptive platform trials in practice - lessons learned

15:30-16:00

Presenter/Author: J. Jack Lee

Affiliations: The University of Texas MD Anderson Cancer Center, Houston

Adaptive platform trials are master-protocol clinical trials designed to evaluate multiple interventions, disease subtypes, treatment combinations, or biomarker-defined cohorts within a single infrastructure, with prospectively planned adaptations based on accumulating data, and it can run perpetually. Prominent features include shared control groups, response-adaptive randomization,

interim decision-making, dropping toxic or ineffective arms, adding new arms, and seamless transition across trial stages. Compared with conventional stand-alone randomized trials, adaptive platform trials can improve efficiency, accelerate learning, reduce duplication of infrastructure, and increase the probability that patients receive promising therapies. However, they also introduce operational, regulatory, statistical, and logistical complexities, including control of type I error, non-concurrent controls, delayed outcomes, multiplicity, changing standards of care, data-sharing governance, and the need for rapid, reliable decision-making based on proper endpoints. Prominent examples of adaptive platform trials include I-SPY 2 and GBM AGILE in oncology, REMAP-CAP and RECOVERY in infectious diseases and critical care, and emerging platform trials in stroke, emergency medicine, etc. Key statistical innovations include Bayesian inferential frameworks, hierarchical modeling, borrowing of information across arms or subgroups, adaptive randomization, predictive probabilities, and model-based interim decision rules. Seamless phase II/III platform trials allow ineffective agents to be eliminated early while promising agents "graduate" for confirmatory evaluation. More recently, adaptive platform concepts have moved earlier in development, including seamless phase I/II and dose-optimization platforms aimed at identifying the optimal biological dose especially for targeted agents and immunotherapies. Despite their promise, successful implementation requires careful planning, extensive simulation of operating characteristics, robust infrastructure, clear governance, timely endpoint ascertainment, and close collaboration among statisticians, clinicians, sponsors, regulators, and data monitoring committees. Practical software tools, including R packages MAMS and OCTOPUS, as well as commercial platforms such as FACTS and East, are increasingly important for design, simulation, monitoring, and reporting of complex adaptive platform trials.

Basket Trials for Efficient Precision Medicine Trials

16:00-16:30

Presenter/Author: James Wason

Affiliations: Newcastle University, Newcastle Upon Tyne

Although rare diseases individually affect a small proportion of people, there are many thousands of them and collectively they cause considerable burden on a high proportion of the population. In parallel, precision medicine is leading to rethinking the definition of some common diseases, potentially reclassifying them into smaller molecularly defined subgroups that may experience quite different treatment effects. It is challenging to conduct clinical trials for rare diseases due to the lower numbers of potential participants and as a result many patients suffering from a rare disease (or category of more common conditions) lack hope of effective treatment. Basket trials, which combine multiple sub-trials testing a common intervention for different related conditions, have a big role to play in making more trials feasible to do for rare diseases. Through a basket trial, it is possible to borrow information on the intervention's effect from one condition when analysing another. In this talk I will present recent methodological developments and outline a research agenda for: - How basket trials can allow jointly evaluating rare and more prevalent conditions with shared biological mechanisms within a unified protocol. - How Bayesian borrowing approaches can be used when the different

conditions have distinct clinical outcomes but a common mechanistic outcome. - Hybrid basket designs combining randomised and external-control subtrials within a single framework. Throughout I will illustrate methods with some real case studies in different disease areas. These examples will highlight both the opportunities and the practical challenges in designing efficient, flexible basket trials for precision medicine and rare disease settings.

Implementing Response-Adaptive Randomization in Rare-Disease Trials: Design Challenges and Exact Solutions

16:30-17:00

Presenter/Author: Sofía S Villar

Affiliations: MRC Biostatistics Unit, University of Cambridge, Cambridge

Response-adaptive randomization (RAR) offers a compelling framework for master protocols by allowing the trial to evolve based on accumulating evidence, potentially directing more patients to more effective treatments. However, implementing these designs in rare disease settings-where patient populations are small and often highly stratified-presents unique challenges that are frequently overlooked in theoretical literature. In rare disease trials, the validity of standard frequentist tests is often compromised because the large-sample (asymptotic) assumptions required for p-value calculations do not hold. This issue is compounded in RAR, where dependence is introduced between outcomes and sample sizes, complicating standard inference. Furthermore, the presence of missing data at interim stages and the need to translate continuous probabilities into practical, clinician-accepted allocation ratios create significant barriers to the use of RAR in master protocols. In this talk, motivated by the phase-II STRATOSPHERE trial-a stratified multi-arm RAR Phase-II trial in a rare respiratory condition-I will present practical and methodological solutions to these hurdles: Mapping Strategies for Desirable Allocations: I will discuss how a 'Mapping' approach can discretize RAR probabilities into fixed allocation ratios, ensuring the design remains faithful to its targets even in very small samples. This approach also bridges the gap for external trial providers by translating RAR probabilities into discrete ratios, it allows the design to be implemented through conventional block-randomization systems already in use by external vendors. Stabilizing RAR via Discretization: I will show that this mapping strategy stabilizes the RAR process by effectively introducing a region where no adaptation occurs. This stabilization is particularly beneficial for the trial's operating characteristics and helps deal with missing or incomplete data at interim analyses. Exact Error Control for RAR Multi-Arm Trials: I will illustrate how alternative approaches to inference (including exact tests and randomization-based inference) can provide a more robust alternative to asymptotic tests for managing Type-I error and multiplicity in rare disease RAR trials. By using the STRATOSPHERE trial as a case study, this presentation will show how exact inference and practical mapping can bridge the gap between RAR theory and its successful implementation in the precision medicine and rare disease landscape.

Tue 29.09.2026

11:00–12:30

AI/ML-Driven Causal Inference with Real-World Data Complexities

Tue, 29.09.26, 11:00–12:30, Room: Rolf Böhme Saal

Chair: Liangyuan Hu

Organizers: Liangyuan Hu

This session advances methods that make modern AI/ML decision-ready for biomedical research where data are censored, heterogeneous, and small in key subgroups.

We showcase approaches that (i) perform causal discovery across multi-cohort data, leveraging overlap while guarding against cohort-specific biases and shifts; (ii) transfer information from large external cohorts to under-represented populations via principled weighting that aligns covariates and outcomes while down-weighting mismatched sources; and (iii) estimate individualized causal pathways with survival outcomes using deep generative models.

Across these contributions, common pillars are explicit estimands, assumption transparency, and valid uncertainty with theory-backed guarantees, paired with scalable computation for longitudinal and high-dimensional settings. The importance is practical and immediate: clinicians and policymakers increasingly rely on AI trained on imperfect data.

By addressing censoring, sample-size limitations, domain shift, and hidden bias, this session delivers tools that convert complex real-world data into credible, transportable causal evidence for precision medicine and public-health decisions.

Speakers

- **Vanessa Didelez**: Leibniz-Institut for Prevention Research und Epidemiology, Germany
- **Yi Li**: University of Michigan, School of Public Health, Department of Biostatistics, USA
- **Xinyuan Song**: Chinese University of Hong Kong, Department of Statistics, Hong Kong, China

---

Improving(?) Causal Discovery with Machine Learning in Practice

11:00-11:30

Presenter/Author: Vanessa Didelez

Affiliations: Leibniz Institute for Prevention Research and Epidemiology - BIPS, Bremen, Germany, Bremen; University of Bremen, Bremen

Causal discovery has its roots in probabilistic approaches to artificial intelligence and has recently seen a great revival. Answering causal questions is the ultimate aim of many scientific enquiries. Much progress has been made over the last decades in formalising causality as a distinct objective of data analyses whenever we are concerned with relations that should remain valid under changes to a system such as interventions, shifts or manipulations. In causal discovery we are concerned with learning the causal structure of a system as a whole (or key parts of it). The structre is typically represented by a causal directed acyclic graph (DAG), so that the statistical task is to find such a DAG (or a set of DAGs or an equivalence class) that fits the data well. Numerous methods for causal discovery have been proposed over the last more than two decades with an explosion of proposals in the last few years. However, despite all of these suggestions, causal discovery is also well-known to not perform very well in practice: the output typically contains many errors / implausibilities and is very unstable. In this talk, I will address a number of practical problems for causal discovery: most methods deal poorly with missing data there is no general method for mixed variable types (discrete / continuous) and any known temporal order should be efficiently taken into account. We have developed some solutions that involve machine learning methodology, e.g. multiple imputation using random forests kernel-based conditional independence tests for mixed variables, where we embed variables in reproducing kernel Hilbert spaces (RKHSs) and replace the usual kernel ridge regression by an RKHS-valued extension of random forests and an interactive 'expert-in-the-loop' approach to incorporate more general background knowledge (in addition to temporal ordering), where the 'expert' could be a large language model. I will discuss these advances and their usefulness in the context of an applied example based in the IDEFICS/I.Family cohort study where we were interested in the causal structure between modifiable behaviours in early life and later childhood health outcomes.

Enhancing Inference for Small Cohorts via Transfer Learning and Weighted Integration of Multiple Datasets

11:30-12:00

Presenter/Author: Yi Li

Affiliations: University of Michigan, School of Public Health, Department of Biostatistics, USA

Lung sepsis remains a significant clinical concern in the Northeastern U.S., yet patients from this region are underrepresented in national databases such as the eICU Collaborative Research Database. This underrepresentation limits reliable inference on key clinical markers, FiO2, creatinine, platelets, and lactate, that reflect oxygenation, kidney function, coagulation, and metabolism, and may differ by sex. To address this challenge, we introduce transfer learning with weights (Translate), a novel adaptive weighting framework for statistical transfer learning across multi-cohort observational data. Translate learns cohort-specific weights that account for both cohort prevalence and

alignment with a small anchor cohort, enabling data-driven, estimand-agnostic integration of information across dissimilar external cohorts. These weights reflect each cohort's effective contribution to an anchor-aligned pseudopopulation and naturally downweight non-comparable cohorts. The method generalizes existing weighting strategies, offers theoretical guarantees for improved precision, and supports a broad class of estimands, including means, variances, distribution functions, covariances, and pairwise correlations. Simulation studies and a real-data application to lung sepsis show that Translate yields more accurate and robust inference for small anchor cohorts by effectively leveraging information from diverse external cohorts.

CGAN for individualized causal mediation analysis with survival outcome.

12:00-12:30

Presenter: Xinyuan Song

Authors: Xinyuan Song [1]; Cheng Huan [1]; Hongwei Yuan [2]

Affiliations: [1] The Chinese University of Hong Kong, Hong Kong; [2] University of Macao, Macao

Causal mediation analysis aims to investigate the mechanism by which an exposure exerts its effects on the outcome, mediated by intermediate variables. However, existing methods for causal mediation analysis in the context of survival models are primarily focused on estimating average causal effects and are difficult to apply to precision medicine. Recently, machine learning has emerged as a promising tool for precisely estimating individualized causal effects without assuming specific model forms. This study proposes a novel method, conditional generative adversarial network (CGAN)-based individualized causal mediation analysis with survival outcomes (CGAN-ICMA-SO), to infer individualized causal effects with survival outcomes based on the CGAN framework. We show that the estimated distribution of the proposed inferential conditional generator converges to the true conditional distribution under mild conditions. Our numerical experiments indicate that CGAN-ICMA-SO surpasses five other state-of-the-art methods. Applying the proposed method to an Alzheimer's disease (AD) Neuroimaging Initiative dataset reveals the individualized direct and indirect effects of the APOE-epsilon4 allele on time to AD onset.

Wed 30.09.2026

09:00–10:30

Accelerating evaluation and adoption of complex digital outcome measures in trials

Wed, 30.09.26, 09:00–10:30, Room: Rolf Böhme Saal

Chair: James Carpenter

Organizers: James Carpenter

The rapid evolution of technological devices in biomedicine allows measuring increasingly complex and high-dimensional data (high-frequency time series, 3D images generated by medical scanners etc). Even with such availability of digital device monitors, relatively few are accepted for primary outcome measurement in late phase trials. This session explores the issues, and how they can be addressed.

Among many possible approaches, functional data analysis methods provide the basic tools to extract knowledge from such intrinsically smooth biomedical data and constitute a promising frontier of research in high-dimensional statistics.

Speakers

- **Valeria Vitelli**: University of Oslo, Institute of Basic Medical Sciences, Oslo Centre for Biostatistics and Epidemiology, Department of Biostatistics, Norway
- **Mia Tackney**: University of Cambridge, MRC Biostatistics Unit, UK
- **Patrick Bastian**: Aarhus University, Denmark

Understanding mortality following elective surgery complications via functional data modeling of patient sedation monitoring data

09:30-10:00

Presenter/Author: Valeria Vitelli

Affiliations: University of Oslo, Oslo

We analyse a large Norwegian cohort of approximately 300 thousands individuals undergoing non-cardiac elective surgery, and study the determinants of deaths occurring up to 1 year after surgery.

Previous studies have found that perioperative complications are associated with mortality, but such studies faced several methodological limitations, did not focus on risk after discharge, and were limited in sample size. Furthermore, such studies have not explicitly linked longer-term deaths to (early postoperative) complications. We develop novel functional data analysis techniques, and particularly multivariate functional clustering and joint latent variable modeling, to the purpose of characterizing the diversity in the sedation patterns and responses under complications, and ultimately understanding the drivers of the deterioration of patients' conditions. When monitoring curves are observed partially, either sparsely or completely latent, the problem can be handled via the family of generalized functional regression models, where the latent response variable is a latent Gaussian Process, and a realization as a sequence of correlated binary observations is observed. We propose novel ways to handle the smoothness constraints on the latent response curves, which allow easily controlling the smoothness level of functional unknowns, and we also propose practical computational frameworks for inference.

Capturing the Curve: Functional Data Analysis for Validated Digital Outcome Measures

09:00-09:30

Presenter: Mia Tackney

Authors: Mia Tackney [1]; Sofía S Villar [2]; James Carpenter [3]

Affiliations: [1] MRC-Biostatistics Unit, University of Cambridge, Cambridge; [2] MRC Biostatistics Unit, University of Cambridge, Cambridge; [3] UCL Innovative CLinical Trial Unit, London

The use of Digital Health Technologies to measure outcomes in clinical trials opens new opportunities as well as methodological challenges. Digital outcome measures can provide more convenient data collection via remote assessment and may potentially detect clinical change earlier. However, there are unresolved methodological challenges around how digital outcomes should be defined and validated, and very few digital outcomes are currently accepted as primary outcomes in late-phase trials. A major difficulty lies in summarizing the volume of data collected by digital devices. Many physiological signals, such as electrocardiogram (ECG) traces or measurements of gait, have a hierarchical functional structure: repeated physiological functions are observed over time and on multiple time scales (seconds, days, weeks). While many summaries can be derived from digital data, typically, only a small subset of pre-defined, clinically interpretable scalars is validated as clinical outcome measures. As an alternative approach, we evaluate data-driven summaries or scores from Multilevel Functional Principal Component Analysis (MFPCA). These scores are derived from a formal method for obtaining meaningful low-dimensional representations of functional data with robust statistical properties. Through a simulation study motivated by smartwatch electrocardiogram (ECG) signals, we compare the performance of MFPCA scores with pre-specified summaries on validation criteria for digital outcome measures. We discuss potential and challenges for this approach in the clinical trials context.

Change point detection in functional data with applications to biomechanical fatigue data

10:00-10:30

Presenter: Patrick Bastian

Authors: Patrick Bastian [1]; Holger Dette [2]

Affiliations: [1] Aarhus University, Aarhus; [2] Ruhr-Universität Bochum, Bochum

Injuries to the lower extremity joints are often debilitating, particularly for professional athletes. Understanding the onset of stressful conditions on these joints is, therefore, important in order to ensure prevention of injuries as well as individualised training for enhanced athletic performance. We study the biomechanical joint angles from the hip, knee and ankle for runners who are experiencing fatigue. The data is cyclic in nature and densely collected by body-worn sensors, which makes it ideal to work with in the functional data analysis (FDA) framework.

We develop new change point detection methods for functional time series, both in the offline and online setting. We improve the state of the art in at least two respects: First, the curves are compared with respect to their maximum absolute deviation, which leads to a better interpretation of local changes in the functional data compared to classical L2-approaches. Second, as slight aberrations are to be often expected in a human movement data, our method will not detect arbitrarily small changes but hunts for relevant changes, where maximum absolute deviation between the curves exceeds a specified threshold, say $\Delta > 0$. This allows us to identify changes associated to fatigue of the runner in a biomechanical data set.

11:00–12:30

Biostatistics for personalized medicine

Wed, 30.09.26, 11:00–12:30, Room: Rolf Böhme Saal

Chair: Antonia Zapf, Maria Grazia Valsecchi, Valeria Vitelli

Organizers: Antonia Zapf, Grazia Valsecchi, Valeria Vitelli

In the era of precision medicine, the paradigm of the "average population trial" is challenged by the need to adopt individualized therapies where treatment choices are dynamically linked to the course of the disease and tailored to the single patient. Biomarker discovery is critical for precision medicine because it reveals molecular features that predict response or risk with targeted treatments.

In this setting, the development of modern trial designs and of computationally intensive statistical methods, e.g. for integrating multi-omics data or drug response profiles, is becoming central for predicting a patient's prognosis or treatment response, and for accelerated drug discovery and companion diagnostics, for instance in oncology.

Speakers

- **Manuela Zucknick**: University of Oslo, Institute of Basic Medical Sciences, Oslo Centre for Biostatistics and Epidemiology, Department of Biostatistics, Norway
- **Eric Laber**: Duke University School of Medicine, Department of Biostatistics & Bioinformatics, USA
- **Linda Marchioro**: Independent statistical contractor; EMA Methodology ESEC Member

---

Leveraging Structured Priors and Multi-Modal Data for Biologically Informed Personalized Medicine

11:30-12:00

Presenter/Author: Manuela Zucknick

Affiliations: Oslo Centre for Biostatistics and Epidemiology, University of Oslo and Oslo University Hospital, Oslo

Introduction

A primary challenge in personalized medicine is the integration of high-dimensional, multi-modal data (including multi-omics, ex vitro drug screens, and clinical information) to tailor treatments and predict survival, often under the constraint of limited sample sizes. Biomedical data are characterised by significant heterogeneity, representing substantial biological variability between individuals as well as strong intra-individual variation, but often also large technical uncertainty, e.g. due to measurement error. Hence, full uncertainty quantification and propagation is essential to delineate biological signal from technical noise.

Methods

We propose a Bayesian hierarchical framework as a unified approach to model these complex biomedical structures, allowing for the tracking of sources of variability and uncertainty. We employ a multivariate Bayesian variable and covariance selection setup to address multi-response settings, improving model performance by leveraging data structures through structured priors, borrowing information across correlated response variables and integrating external biological knowledge. Key extensions include the use of Markov random field priors for latent variable selection to exploit molecular pathways and the incorporation of interaction effects with moderating variables to account for heterogeneity in omics impacts. Results and Discussion We demonstrate some of the modelling aspects in applications in precision medicine and cancer drug screens. Overall, this modelling approach facilitates more precise individual predictions, even when sample sizes are small, by combining diverse data modalities in a biologically informed way and by borrowing information across multiple related outcomes. The joint analysis framework ensures that clinical interventions are informed by a comprehensive, uncertainty-aware integration of patient-specific data and existing biological knowledge.

Non-dominated online treatment strategies

12:00-12:30

Presenter/Author: Eric Laber

Affiliations: Duke University, Durham

Sequential experiments are often designed to balance the generation information and utility. A sequential experimental design that maximizes long-term cumulative utility must explore, i.e., forgo immediate utility in favor of information, only if and when the value of information generated by exploration outweighs the short-term loss in utility. Unfortunately, it is not possible to construct such designs in general. Instead, researchers rely on heuristics that force exploration through randomization or by adding an exploration "bonus" to the utility function. We show that experiments selected by commonly used heuristics frequently generate both less information and less utility than alternatives. Such experiments are said to be dominated in terms of utility and information. We show that if one removes from consideration experiments that are estimated to be dominated at each time point and then applies standard heuristics for forced exploration, it is still possible to obtain consistency and asymptotic normality for maximum quasi likelihood estimators (MQLE). Furthermore, in a suite

of simulation experiments, we show that these non-dominated heuristics generate more utility and more information (i.e., are more efficient) than their standard counterparts.

Predictive biomarker assays in the context of medicinal product lifecycle: The journey towards a European Guideline

11:00-11:30

Presenter/Author: Linda Marchioro

Affiliations: Independent statistical contractor; EMA Methodology ESEC Member

Introduction

In the framework of personalised medicine, an increasing number of medicinal products is being investigated or authorised in biomarker-defined subpopulations. Identifying patients most likely to benefit from treatment, or those at increased risk of adverse reactions, relies on predictive biomarker assays. This creates an interface with the medical device legislation, notably Regulation (EU) 2017/746 on in vitro diagnostic medical devices. The European Medicines Agency (EMA) is developing a guideline to address medicinal product requirements in the context of co-developed medicinal products and their associated assays (companion diagnostics, CDx), as well as related situations such as follow-on CDx (i.e. diagnostic assays developed after the initial medicinal product authorisation, but with the same clinical indication and intended purpose as a predecessor CDx).

Methods

The guideline has been developed by a drafting group of experienced assessors from the EMA network and is currently undergoing consultations (internal and public) to incorporate stakeholders' feedback.

Results

The guideline provides recommendations on key aspects, including the scientific rationale for the biomarker, the analytical validation of the biomarker assay, the clinical validation of both the biomarker and the assay, and relevant regulatory considerations (including the inclusion of information in the Summary of Product Characteristics, SmPC). Statistical aspects include approaches to evidence generation depending on the development phase and prior rationale on the biomarker, the definition of cut-offs for continuous biomarkers, and requirements for bridging studies (i.e. when the assay and the medicinal product are not co-developed). Conclusions This guideline aims to support developers and regulators in the development of safe and effective medicines within a personalised medicine framework, while clarifying the respective roles and boundaries of medicinal product and in vitro diagnostic legislations.

14:00–15:30

Evaluating Fairness in Clinical Prediction Models and AI: Current and Emerging Topics

Wed, 30.09.26, 14:00–15:30, Room: Rolf Böhme Saal

Chair: Laure Wynants, Ben Van Calster

Organizers: Laure Wynants, Ben van Calster

AI fairness is an important topic for AI policy, and for research and healthcare. The recently published TRIPOD+AI reporting guideline and the TRIPOD AI risk of bias tool mention fairness, but do not specify how to evaluate fairness. This session starts with a discussion of the results of a systematic review of clinical predictive AI fairness metrics (https://arxiv.org/abs/2506.17035), that unveiled important shortcomings and a fragmented landscape.

Next, we investigate assumptions underlying decision curve analysis that have implications for fairness evaluation. Lastly, we present a software solution to empower independent fairness evaluation while preserving intellectual property and viability of prediction models in the context of the evolving regulatory landscape.

Speakers

- **Gary Collins**: University of Birmingham, Department of Applied Health Sciences, UK
- **Joie Ensor**: University of Birmingham, Department of Applied Health Sciences, UK
- **Junfeng Wang**: Maastricht University, Care and Public Health Research Institute, Department of Epidemiology, Netherlands

---

A Fractured Landscape: Evaluating Fairness in Clinical Prediction Models

14:00-14:30

Presenter/Author: Gary Collins

Affiliations: Department of Health Sciences, School of Health Sciences, college of Medicine and Health, University of Birmingham, Birmingham, United Kingdom, Birmingham

Clinical prediction models have the potential to improve patient outcomes, but only if fairness is adequately measured and addressed. In the context of clinical prediction models, fairness refers to the property whereby individuals or groups defined by sensitive attributes, such as age, sex, race

or ethnicity, or socioeconomic status, are not systematically disadvantaged by model outputs or resulting clinical decisions. In this talk I will provide a critical evaluation of the current fairness evaluation landscape. Using a taxonomy classifying metrics by performance dependency, model output level, and base performance metric, I will expose a fragmented landscape dominated by threshold-dependent metrics sensitive to arbitrary decision cutoffs. A further underappreciated methodological challenge is metric non-collapsibility: several widely used performance metrics, including the c-statistic, calibration intercept, and calibration slope, cannot be expressed as weighted averages of subgroup-specific estimates. This means apparent differences in fairness across subgroups may reflect case-mix variation rather than genuine disparities, potentially misleading fairness evaluations. Very few metrics were developed for healthcare, reflecting a disconnect between methodological development and clinical applicability. Gaps in intersectional analysis and real-world validation further limit existing approaches. I will discuss implications for TRIPOD+AI, FUTURE-AI, and STANDING Together, which all address fairness with caution but limited specificity. I will argue that meaningful progress requires these guidelines to explicitly acknowledge key methodological challenges in fairness evaluation.

When Net Benefit Misleads the Decision: A Cautionary Note on Population-Level Decision Curve Analysis in the Presence of Subgroups

14:30-15:00

Presenter: Junfeng Wang

Authors: Junfeng Wang [1]; Kim Zhipei Wang [1]; Ben Van Calster [2]; Nan van Geloven [3]; Ewout Steyerberg [4]; Laure Wynants [1, 2]

Affiliations: [1] Maastricht University, Maastricht; [2] KU Leuven, Leuven; [3] Leiden University Medical Center, Leiden; [4] University Medical Center Utrecht, Utrecht University, Utrecht

Introduction

Background Decision Curve Analysis (DCA) is widely used to evaluate prediction models and guide clinical adoption decisions using net benefit (NB). In practice, population-level DCA is often treated as a direct proxy for population utility. However, this interpretation relies on assumptions that are rarely examined. We aimed to formally characterize when population-level DCA may lead to incorrect population-level decisions in the presence of subgroup differences.

Methods

We considered the decision to adopt a prediction model versus default strategies in a population. We analytically examined the relationship between population-level ΔUtility and ΔNB when the population consists of subgroups. We derived the "inconsistency region", defined as the combination of subgroup ΔNB values for which population NB and population utility lead to opposite conclusions. We also developed a practical framework to investigate the impact of subgroup differences in DCA.

Results

We demonstrate that conflicting subgroup-specific NB results serve as an empirical warning signal that population-level NB may recommend model adoption even when overall population utility is not improved. The inconsistency region increases when subgroup sizes are more similar and true positive compared to a false negative (a-c) values more dissimilar. When a-c varies across subgroups, NB implicitly aggregates quantities measured on different scales, violating the conditions under which it is decision-theoretically valid. A real-world example illustrates the practical relevance of this problem. Conclusions Population-level DCA assumes homogeneous a-c values. When this assumption cannot confidently be made, routine subgroup DCA should become standard practice. Our framework can be used to determine if differences in observed subgroup results invalidate population-level decision making.

---

evaluatr: a software system to empower independent fairness evaluation of clinical prediction models

15:00-15:30

Presenter: Joie Ensor

Authors: Joie Ensor [1]; Ben Van Calster [2]; Lasai Barreñada [2]; Laure Wynants [3]; Benjamin Perry [1]

Affiliations: [1] University of Birmingham, Birmingham; [2] KU Leuven, Leuven; [3] Maastricht University, Maastricht

Independent evaluation of clinical prediction models (CPMs) in new populations and settings is fundamental to fairness assessment. Understanding whether a CPM performs equitably across ethnic groups, geographic settings, or socioeconomic strata requires researchers from varied communities to conduct evaluations autonomously. Yet in practice, independent evaluation of promising CPMs remains rare. A key structural barrier is the tension between the open science requirement to publish model specifications and the commercial and regulatory realities facing developers, particularly as regulatory frameworks increasingly classify CPMs as medical devices. This tension is not merely inconvenient, but actively limits which populations get evaluated, and by whom, with predictable consequences for fairness. We present evaluatr, a free R package implementing a two-component system for independent CPM evaluation that resolves this tension. Developers register their CPM with the system and store parameter values securely evaluators use a minimal R workflow (requiring no data-sharing agreement, no specialist infrastructure, and no access to model coefficients) to compute predictions locally on their own data and derive performance metrics of interest. Patient data never leaves the evaluator's environment. Model specifications are never disclosed. The system is directly compatible with the TRIPOD+AI reporting guideline's provision for software-mediated access in place of coefficient disclosure. Crucially, the system removes practical barriers that currently concentrate evaluation in well-resourced settings. Any researcher with their own dataset and a basic familiarity with R can evaluate a CPM, including across subgroups, without negotiating access to the model or sharing sensitive data. We argue this architecture meaningfully expands the evaluator

pool, enabling evaluation in the populations and settings currently most underrepresented and most in need of evidence.

Signature Session: AI and Me

Tue, 29.09.26, 13:30–15:00, Room: Rolf Böhme Saal

Chair: Harald Binder

How does musical improvisation take shape when humans and AI listen, respond and make musical decisions together? In this performative research format, double bassist Jonas Gerigk improvises together with SOMAX2, an AI system for musical interaction developed at IRCAM. SOMAX2 analyzes the musical performance in real time, identifies musical connections within previously learned sonic material and generates its own musical responses. The performance opens an improvised dialogue between human and AI, inviting the audience to experience their musical relationships. It is followed by an open discussion in which questions and observations from the audience are warmly encouraged. The exchange flows directly into a final improvised performance, where the shared reflections become the starting point for new musical decisions.

The artistic research investigates how a personal musical vocabulary is cultivated through long-term improvisational practice. One case study explores this line of inquiry through musical interaction between humans and artificial intelligence and asks how improvised encounters with artificial intelligence can open new perspectives on musical listening, decision-making and practice.

www.jonasgerigk.de

Workshops

Mon 28.09.2026

11:00–12:30

Designing Next-Generation Respiratory Virus Trials: Estimands, Core Outcomes, and Adaptive Pandemic-Ready Frameworks

Mon, 28.09.26, 11:00–12:30, Room: K1

Chair: Inge Christoffer Olsen

Authors: Inge Christoffer Olsen [1]; Franz Koenig [2]; Alain Amstutz [3]; Matthias Briel [3]; Dominique Costagliola [4]; Thomas Jaki [5]; Declan Devane [6]; Johannes Schwenke [7]

Affiliations: [1] Oslo University Hospital, Oslo; [2] Medical University of Vienna, Vienna; [3] University of Basel, Basel; [4] Institut Pierre Louis d'Epidémiologie et de Santé Publique, Sorbonne Université, INSERM, Paris; [5] University of Cambridge, Cambridge; [6] University of Galway, Galway; [7] Department of Clinical Research, CLEAR-Methods Centre, University Hospital Basel and University of Basel, Basel

General outline and objectives

This workshop explores methodological, operational and regulatory challenges in designing clinical studies for respiratory tract viral infections, including pandemic settings. Participants will examine how estimand selection, core outcome set development, and adaptive protocol frameworks influence trial interpretability, comparability, and decision-making. Special attention will be given to the emerging role of AI when designing a trial, e.g., using AI in the work of identifying core outcome sets. Furthermore, we will elaborate on how to deal with operational, legal and logistical hurdles in case of health emergency situations and how to potentially overcome them, e.g., implementing a federated trial approach to start as quickly as possible in different indications. Finally, we will look at choice of development phase and how to select candidate interventions.

The workshop aims to provide practical guidance for selecting meaningful outcomes, designing flexible trial infrastructures (umbrella, master, or pathogen-specific protocols), and aligning designs with regulatory expectations. Using real case studies from large European pandemic research initiatives, participants will learn how to translate methodological principles into operational study design.

Learning objectives:

Understand options and implications of estimand selection in respiratory virus trials Evaluate alternative platform and pandemic-ready trial designs Learn structured approaches to developing core

outcome sets, including AI-assisted methods Assessing clinical trial design by the use of clinical trial simulations Understand issues in choosing the appropriate interventions and phase of development

Structure

11:00 - 11:10 Inge Christoffer Olsen: Setting the scene: challenges in respiratory viral infection research

11:10 - 11:25 Thomas Jaki: RECOVERY - an adaptive platform trial during the COVID-19 pandemic

11:25 - 11:40 Inge Christoffer Olsen: Trial infrastructure: umbrella, platform, and pathogen-specific designs for pandemic preparedness

11:40 - 11:55 Julia Louw: Core outcome sets: development process and lessons from AI-supported methods

11:55 - 12:05 Johannes Schwenke: Simulation Tools for Adaptive Viral Respiratory Disease Trials

12:05 - 12:30 All: Panel discussion and key recommendations

Setting the scene: challenges in respiratory viral infection research

11:00-11:10

Presenter/Author: Inge Christoffer Olsen

Affiliations: Oslo University Hospital, Oslo

We are living in an ever more globalized world where infectious diseases have the potential to spread fast and wide. This was made very clear in the COVID-19 pandemic, but also in the recent mpox, hantavirus and ebola outbreaks. One of the pillars of countermeasures is therapeutic treatments but we need clinical trials to provide rapid evidence. In this presentation I will talk of the challenges we face as methodologist, and how we might meet them. The talk will also set the stage for the other talks in this workshop.

RECOVERY - an adaptive platform trial during the COVID-19 pandemic

11:10-11:25

Presenter/Author: Thomas Jaki

Affiliations: University of Cambridge, Cambridge

RECOVERY is a randomised, controlled, open-label platform trial comparing several possible treatments with usual care in patients admitted to hospital with COVID-19. It was set up in record time

to seek treatment options for the newly identified SARS-CoV-2 virus in early 2020. In this presentation, I provide an overview of the design of the study and provide input into some of the discussions around some design aspects with a focus on endpoint choice.

Trial infrastructure: umbrella, platform, and pathogen-specific designs for pandemic preparedness

11:25-11:40

Presenter/Author: Inge Christoffer Olsen

Affiliations: Oslo University Hospital, Oslo

How do we plan for the next pandemic? What have we learned from the COVID-19 pandemic, and how can we use this knowledge to do better next time? I will present the PROACT EU-Response project and the EU-SyndAct umbrella protocol, and use these examples to discuss around how we can prepare and what we need to prepare. I will focus on the methodology, but will also touch upon other aspects.

Core outcome sets: development process and lessons from AI-supported methods

11:40-11:55

Presenter: Julia Louw, Declan Devane

Authors: Declan Devane

Affiliations: University of Galway, Galway

Inconsistent outcome selection across trials of acute viral respiratory infections (AVRIs) limits evidence synthesis and slows decision-making during health emergencies. Existing core outcome sets (COS) focus on COVID-19 and contain pathogen-specific elements that constrain broader applicability. COSAVRI is developing an internationally agreed COS for pharmacological treatments in hospitalised adults with AVRIs, following the COMET Handbook and COS-STAD recommendations, and registered with COMET (No. 3486) and OSF. This presentation, on behalf of the COSAVRI Steering Group, walks attendees through the four-stage COSAVRI process and focuses on the LLM-assisted, human-checked pipeline used to consolidate outcomes between Stage 1 (rapid scoping review of RCTs published 2015 to 2025) and Stage 2 (Real-Time Delphi survey). Across 360 unique RCT reports we extracted 3,237 raw outcome records. Claude Opus 4.6 within Claude Cowork supported terminology standardisation, deduplication, semantic clustering, composite outcome review, eligibility screening and Delphi curation ChatGPT Pro 5.4 was used as an independent check during extraction validation. All outputs were treated as proposals where each was checked, edited or rejected by the research team, and two clinicians independently reviewed the eligible outcome list and completed a final sense-check of the Delphi candidates. The pipeline produced 2,287 unique records,

597 clusters, 251 eligible outcomes, 234 post-clinical-review outcomes, and 185 candidate Delphi items, with a version-controlled audit trail of approximately 2,195 cluster-level decisions. Validation against human extraction of 60 reports (about 17% of the cleaned set), including a held-out final set of 30 reports, gave a micro-averaged F1 of 0.982 245 of 252 reference outcomes were captured, two additional valid outcomes were identified and seven were missed.

Simulation Tools for Adaptive Viral Respiratory Disease Trials

11:55-12:05

Presenter/Author: Johannes Schwenke

Affiliations: Department of Clinical Research, CLEAR-Methods Centre, University Hospital Basel and University of Basel, Basel

EU-SyndAct is an emerging adaptive clinical trial platform for acute viral respiratory infections. For EU-SyndAct, the planned primary analysis uses a Markov ordinal state transition model. In this presentation, we will introduce open-source software developed to support trial simulation and design for EU-SyndAct. The package facilitates simulation of longitudinal ordinal patient data, preparation of data for Markov ordinal state transition models, estimation of state occupancy probabilities and their uncertainty, and visualization of results. We will also discuss how this workflow may be connected to other modular platform-trial simulation tools, allowing for simulations of operating-characteristic evaluations for adaptive trial designs.

From Research Question to Access: Using Nationwide Claims Data Today - and Electronic Health Record Data Tomorrow

Mon, 28.09.26, 11:00–12:30, Room: K4

Chair: Steven Wolter

Authors: Steven Wolter

Affiliations: Forschungsdatenzentrum Gesundheit, Bundesinstitut für Arzneimittel und Medizinprodukte, Bonn

General outline and objectives

The Forschungsdatenzentrum Gesundheit (FDZ Gesundheit / Health Data Lab at the German Federal Institute for Drugs and Medical Devices, BfArM) provides controlled access to nationwide statutory health insurance (SHI) claims data covering approximately 74 million insured persons in Germany. These longitudinal population-level data enable large-scale analyses in epidemiology, biostatistics, health services research, and pharmacoepidemiology.

The FDZ data infrastructure is currently evolving in two important ways. First, a new Data Model 4 will soon be introduced, expanding the scope and structure of available claims information and thereby enhancing analytical possibilities. Second, the FDZ is expanding its portfolio to include electronic health record (EHR) data. Starting at the end of 2026, the first EHR data type - the electronic medication list from the German electronic health record - will be made available for scientific use. In the future, researchers will be able to combine claims data and EHR-based information, opening new perspectives for real-world evidence generation and medication-related research.

This interactive workshop will guide participants through the pathway from research idea to data access at the FDZ, with a focus on both current analytical opportunities and the evolving data landscape. The objectives are to:

Introduce the FDZ data infrastructure, including the current claims data, the upcoming Data Model 4, and the integration of EHR medication data, outlining their respective strengths and limitations. Demonstrate which research questions can be addressed today, how analytical potential expands with Data Model 4, and how study concepts may evolve with the availability of EHR data. Provide practical guidance on submitting a high-quality data use application, including best practices, common pitfalls, and strategies to prepare projects in anticipation of future data extensions.

Structure

Presentation: Intro to the FDZ Gesundheit and data available for use

Moderated Q&A

Presentation: Explore FDZ Data

Interactive: What questions can be answered with the FDZ data?

Presentation: How to prepare

Interactive: Download Public Use File and code example questions

---

Getting Started with the German Health Data Lab: Data Resources, Access Pathways, and Research Applications

11:00-12:30

Presenter: Steven Wolter, Hannah Pauly

Authors: Steven Wolter

Affiliations: Forschungsdatenzentrum Gesundheit, Bundesinstitut für Arzneimittel und Medizinprodukte, Bonn

The Forschungsdatenzentrum Gesundheit (FDZ Gesundheit, Health Data Lab at the German Federal Institute for Drugs and Medical Devices, BfArM) provides secure access to nationwide statutory health insurance (SHI) claims data covering approximately 74 million insured persons in Germany. These longitudinal population-level data support a broad range of research, including epidemiology, health services research, pharmacoepidemiology, and health economics.

This introductory talk provides the foundation for the subsequent workshop sessions by familiarizing participants with the FDZ infrastructure, its data resources, and the process of obtaining access for scientific research. It presents the current scope of available claims data, introduces the forthcoming Data Model 4, and provides an outlook on the planned integration of electronic health record (EHR) data, beginning with the electronic medication list. Together, these developments will substantially expand opportunities for generating real-world evidence and addressing increasingly complex research questions.

In addition to the available data, the presentation explains the pathway from a research idea to a successful data use application. Participants will gain an overview of the application process, eligibility requirements, documentation, review procedures, and key considerations when translating a scientific question into a feasible study proposal. Common challenges and frequent reasons for revision will be discussed to support efficient project planning.

The talk establishes a shared understanding of the FDZ data landscape and prepares participants for the interactive workshop exercises, in which they will explore research feasibility, develop study concepts, and align research questions with current and future data availability. It is intended for researchers seeking to conduct high-quality real-world evidence studies using German routine healthcare data.

Developing and Executing Privacy-Preserving Federated Analyses with FLAME

Mon, 28.09.26, 11:00–12:30, Room: N3

Chair: Marius Herr

Authors: Marius Herr [1]; Peter Placzek [2]; Bruce Schultz [3]; David Hieber [2]; Alexander Röhl [2]; Mehrshad Jaberansary [4]; Hammam Abu Attieh [5]; Mehmed Halilovic [6]; Toralf Kirsten [7]; Fabian Prasser [8]; Oliver Kohlbacher [2]

Affiliations: [1] Institute for Translational Bioinformatics, University Hospital Tübingen, Tübingen; [2] Interfaculty Institute for Biomedical Informatics (IBMI), University Tübingen, Tübingen; [3] Institute for Biomedical Informatics (BI-K) - University of Cologne, Köln; [4] Institute for Medical Informatics, University of Cologne, Cologne; [5] Medical Informatics Group, Center of Health Data Science, Berlin Institute of Health at Charité - Universitätsmedizin Berlin, Berlin; [6] Berlin Institute of Health @ Charité, Berlin; [7] Leipzig University Medical Center, Dept. Medical Data Science, Leipzig; [8] Medical Informatics Group, Berlin Institute of Health at Charité - Universitätsmedizin Berlin, Berlin

General outline and objectives

Data protection regulations such as GDPR require that the provision of data in healthcare balances strict protection of patient data with the societal benefits of medical research. At the same time, institutional data silos hinder cross-site analytics and collaborative model development. Federated learning (FL) enables distributed analyses without transferring sensitive data, but practical adoption depends on usable tooling, transparent governance, and reproducible execution of workflows.

This workshop introduces FLAME, an open-source federated analytics platform developed within the national PrivateAIM project, and focuses on the practical lifecycle of developing, submitting, and executing analyses using a public demonstration instance and FLAME software development kit (SDK). FLAME follows a hub-and-node architecture in which analyses are executed locally at data-providing institutions while coordination and aggregation are managed centrally. The SDK abstracts communication, data access, and result handling, enabling developers to focus on methodological implementation. Multiple federated analysis patterns and execution modes are supported and reproducibly deployable across sites.

A wide range of analytical use cases, including imaging-based workflows, privacy-preserving genome-wide association studies, and newly developed federated methods, have already been successfully executed on FLAME. Documented example workflows demonstrate reproducible end-to-end execution across heterogeneous data modalities. The platform is currently transitioning toward productive operation across institutions, accompanied by weekly community meetings for knowledge transfer and feature dissemination.

The workshop is designed as a hands-on interactive session in which participants experience the full federated analysis workflow from local development to distributed execution and result inspection. This practical format addresses current barriers such as the complexity of tooling and uncertainty in cross-institutional collaboration. The session targets researchers, clinicians, data scientists, and

infrastructure developers interested in privacy-preserving medical analytics and aims to foster interdisciplinary understanding and community adoption.

Structure

11:00 - 11:20 Marius Herr: FLAME: Enabling Privacy-Preserving Federated Learning for Sensitive Medical Data

11:20 - 11:50 Marlena Meyer, Marius Herr: Analyses development using the FLAME SDK and demo analysis

11:50 - 12:20 Marlena Meyer, Marius Herr: Analyses submission and execution: The FLAME Hub and Node

12:20 - 12:30

Discussion and Feedback to FLAME

FLAME: Enabling Privacy-Preserving Federated Learning for Sensitive Medical Data

11:00-11:20

Presenter: Marius Herr, Peter Placzek

Authors: Marius Herr [1]; Peter Placzek [1]; Bruce Schultz [2]; David Hieber [1]; Alexander Röhl [1]; Mehrshad Jaberansary [3]; Hammam Abu Attieh [4]; Mehmed Halilovic [5]; Fabian Prasser [6]; Toralf Kirsten [7]; Oliver Kohlbacher [1]

Affiliations: [1] Interfaculty Institute for Biomedical Informatics (IBMI), University Tübingen, Tübingen; [2] Institute for Biomedical Informatics (BI-K) - University of Cologne, Köln; [3] Institute for Medical Informatics, University of Cologne, Cologne; [4] Medical Informatics Group, Center of Health Data Science, Berlin Institute of Health at Charité - Universitätsmedizin Berlin, Berlin; [5] Berlin Institute of Health @ Charité, Berlin; [6] Medical Informatics Group, Berlin Institute of Health at Charité - Universitätsmedizin Berlin, Berlin; [7] Leipzig University Medical Center, Dept. Medical Data Science, Leipzig

The growing demand for collaborative medical research is constrained by strict data protection regulations, institutional data silos, and limited interoperability across healthcare infrastructures. The PrivateAIM project addresses these challenges by providing an open-source federated analytics platform (FLAME) that enables distributed analysis and machine learning without transferring sensitive patient-level data between institutions. The project aims to establish a secure, scalable, and reproducible infrastructure for privacy-preserving medical research across organizational boundaries. FLAME uses a hub-and-node architecture in which analytical workflows run locally at participating institutions, while orchestration, coordination, and aggregation are managed centrally via a federated hub. This architecture lets institutions retain full control over their data while participating in collaborative analyses and model development. The platform integrates multiple privacy-enhancing technologies (PETs) and security mechanisms to support compliant federated analytics across heterogeneous infrastructures. These include isolated execution environments, controlled workflow

deployment, encrypted communication channels, role-based access control, and auditability of analysis execution. In addition, FLAME supports federated execution patterns that minimize exposure of intermediate results and reduce the risk of unintended information leakage.

Analyses development using the FLAME SDK and demo analysis

11:20-11:50

Presenter: Marius Herr, Marlena Meyer

Authors: Marius Herr [1]; Peter Placzek [2]; Bruce Schultz [3]; David Hieber [2]; Alexander Röhl [2]; Mehrshad Jaberansary [4]; Hammam Abu Attieh [5]; Mehmed Halilovic [6]; Fabian Prasser [7]; Toralf Kirsten [8]; Oliver Kohlbacher [2]

Affiliations: [1] Institute for Translational Bioinformatics, University Hospital Tübingen, Tübingen; [2] Interfaculty Institute for Biomedical Informatics (IBMI), University Tübingen, Tübingen; [3] Institute for Biomedical Informatics (BI-K) - University of Cologne, Köln; [4] Institute for Medical Informatics, University of Cologne, Cologne; [5] Medical Informatics Group, Center of Health Data Science, Berlin Institute of Health at Charité - Universitätsmedizin Berlin, Berlin; [6] Berlin Institute of Health @ Charité, Berlin; [7] Medical Informatics Group, Berlin Institute of Health at Charité - Universitätsmedizin Berlin, Berlin; [8] Leipzig University Medical Center, Dept. Medical Data Science, Leipzig

The development and deployment of federated analyses remain a major barrier to adopting privacy-preserving analytics in healthcare research. Distributed execution environments, heterogeneous infrastructures, and complex orchestration mechanisms often require substantial engineering effort beyond the analytical methodology itself. The FLAME software development kit (SDK) addresses these challenges by providing a standardized framework for implementing, executing, and reproducing federated analytical workflows. The SDK abstracts infrastructure-specific functionality, including communication, execution handling, local data access, and aggregation logic, enabling developers to focus on methodological implementation rather than distributed systems engineering. Multiple federated execution patterns are supported and can be reproducibly deployed across heterogeneous environments. Security and governance mechanisms are integrated into the workflow lifecycle through isolated execution environments, encrypted communication, controlled deployment procedures, and institutional approval processes, ensuring that patient-level data remains within local infrastructures. To support onboarding and practical adoption, publicly available demo analyses demonstrate the complete federated workflow from local development to distributed execution and centralized result aggregation. Existing examples include imaging-based workflows, federated differential gene expression analyses, and privacy-preserving genome-wide association studies. These workflows illustrate how analyses can be reproducibly executed across multiple institutions without transferring sensitive data. The workshop also includes a hands-on component in which participants develop and execute their own analysis code using the FLAME SDK. Participants will learn to structure federated workflows, implement local computation and aggregation steps, package analyses for distributed execution, and inspect aggregated results within a federated environment. This

practical introduction aims to lower technical barriers to federated analytics development and foster reproducible, interoperable, and privacy-preserving biomedical research workflows.

Analyses submission and execution: The FLAME Hub and Node

11:50-12:20

Presenter: Marius Herr, Peter Placzek, Marlena Meyer

Authors: Marius Herr [1]; Peter Placzek [2]; Bruce Schultz [3]; David Hieber [2]; Alexander Röhl [2]; Mehrshad Jaberansary [4]; Hammam Abu Attieh [5]; Mehmed Halilovic [6]; Fabian Prasser [7]; Toralf Kirsten [8]; Oliver Kohlbacher [2]

Affiliations: [1] Institute for Translational Bioinformatics, University Hospital Tübingen, Tübingen; [2] Interfaculty Institute for Biomedical Informatics (IBMI), University Tübingen, Tübingen; [3] Institute for Biomedical Informatics (BI-K) - University of Cologne, Köln; [4] Institute for Medical Informatics, University of Cologne, Cologne; [5] Medical Informatics Group, Center of Health Data Science, Berlin Institute of Health at Charité - Universitätsmedizin Berlin, Berlin; [6] Berlin Institute of Health @ Charité, Berlin; [7] Medical Informatics Group, Berlin Institute of Health at Charité - Universitätsmedizin Berlin, Berlin; [8] Leipzig University Medical Center, Dept. Medical Data Science, Leipzig

This workshop focuses on the practical execution lifecycle of federated analyses within FLAME and provides a hands-on introduction to deploying and orchestrating distributed analytical workflows in privacy-preserving research environments. Participants will execute preconfigured analyses in a public demonstration environment and experience the complete workflow from submission and node selection to distributed execution, aggregation, and result retrieval. Analyses are executed locally at participating institutions within isolated containerized environments, and only aggregated or privacy-enhanced intermediate results are exchanged. Raw patient-level data remains entirely within institutional boundaries throughout execution. The workshop further demonstrates operational aspects such as execution monitoring, workflow orchestration, logging, and failure handling across distributed infrastructures. Integrated security and governance mechanisms include encrypted communication, controlled deployment procedures, institutional approval workflows, and isolated execution environments. Example executions include federated machine learning, statistical analyses, and privacy-preserving biomedical workflows, all reproducibly deployed across multiple nodes. By focusing on practical execution workflows, the workshop aims to lower operational barriers to adopting federated analytics and provide participants with hands-on experience in secure, reproducible distributed biomedical analysis.

From Theory to Practice in the EU HTA JCA: Early Insights and Lessons Learned

Mon, 28.09.26, 11:00–12:30, Room: N5

Chair: Kirsten Herrmann

Authors: Kirsten Herrmann [1]; Saskia Trescher [2]; Peter Schlattmann [3]; Beate Wieseler [4]; Stefanie Wüstner [5]; Matthias Wilken [6]

Affiliations: [1] Pharming Group N.V., München; [2] Pfizer Pharma GmbH, Berlin; [3] Universitätsklinikum Jena, Jena; [4] Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG), Köln; [5] AMS Advanced Medical Service, München; [6] Bundesverband der pharmazeutischen Industrie, Berlin

General outline and objectives

This workshop examines how the EU's new Joint Clinical Assessment (JCA) process is transitioning from theory into practice under the recent EU HTA regulation, highlighting early insights and lessons learned with a special focus on the German experience. It will provide a comprehensive overview of the challenges, solutions, and real-world experiences encountered during the first wave of JCA implementations, particularly looking at how joint assessments interface with national procedures like Germany's AMNOG and how methodological differences between EU-level and domestic HTA requirements are being managed.

Structure

11:00 - 11:25 Beate Wieseler: EU HTA Joint Clinical Assessments: Early Implementation Experience from the IQWiG Perspective

11:25 - 11:50 Stephanie Wüstner: Early Lessons from the first HTA procedures assessed at both European JCA and German AMNOG level

11:50 - 12:15 Matthias Wilken: Implementing EU HTA JCAs: Early Lessons and Implications for Pharmaceutical Companies

12:15 - 12:30

Moderated panel discussion integrating insights from all speakers and inviting questions and experiences from the audience.

EU HTA Joint Clinical Assessments: Early Implementation Experience from the IQWiG Perspective

11:00-11:25

Presenter/Author: Beate Wieseler

Affiliations: Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG), Köln

IQWiG is involved in the European JCA procedure in several capacities: as a member of the JCA Subgroup and the HTA Coordination Group, and as an assessor in a number of JCAs. This presentation will address the implementation phase of the EU HTA JCA from these various perspectives, reflecting on the experience and outcome of the initial JCAs. After conclusion of the European process, JCAs are meant to support national HTA, i. e. the AMNOG procedure in Germany. The presentation will discuss the interface of the European and German procedure including first practical experiences.

Early Lessons from the first HTA procedures assessed at both European JCA and German AMNOG level

11:25-11:50

Presenter/Author: Stefanie Wüstner

The first procedures assessed both within the EU Joint Clinical Assessment (JCA) framework and the German AMNOG process provide an early opportunity to examine how European and national requirements for health technology assessment (HTA) interact in practice. These initial experiences are particularly informative because they highlight the interface from two directions: the JCA perspective on how a common evidence package is developed to address heterogeneous Member State needs, and the national perspective on how JCA outputs can be used within AMNOG. This presentation analyses early procedural and methodological observations from these first parallel experiences. It focuses on how assessment scope, analytical requirements, evidence generation, and dossier planning translate across the European and national levels, and how this affects collaboration between global and national market access teams. Three areas will be addressed. First, procedural interfaces, including timelines, dossier finalization, completeness checks, and late adaptations that require coordination between JCA and AMNOG workstreams. Second, methodological interfaces, including comparator selection, outcome operationalization, subgroup and sensitivity analyses, multiplicity-related reporting, and the role of indirect or complementary evidence become that become relevant when translating European evidence packages into the German context. Third, implementation aspects, including which elements of an AMNOG "delta" dossier can be developed in close alignment with the JCA dossier, which require country-specific adaptation, and where structured late-stage cross-checks are critical. Overall, the presentation summarizes early practical lessons from the first procedures across both systems and highlights recurring interface challenges relevant for future submissions.

Implementing EU HTA JCAs: Early Lessons and Implications for Pharmaceutical Companies

11:50-12:15

Presenter/Author: Matthias Wilken

Affiliations: Bundesverband der pharmazeutischen Industrie, Berlin

With the EU Health Technology Assessment Regulation now moving from concept to implementation, Joint Clinical Assessments are becoming a central element of the European market access landscape. For pharmaceutical companies, the JCA is not merely an additional procedural step, but a fundamental change in how clinical evidence must be planned, generated, presented and translated into national reimbursement contexts. From the perspective of the German pharmaceutical industry, early lessons point to both opportunities and challenges. The JCA may reduce duplication and support more consistent scientific assessment across Member States. At the same time, companies face demanding timelines, complex PICO requirements, limited opportunities for iterative exchange, and the need to align European evidence expectations with established national processes such as the German AMNOG benefit assessment. Key practical implications include the need for earlier and more integrated evidence planning across regulatory, HTA, medical, market access and biostatistics functions systematic anticipation of divergent comparator and subgroup expectations and robust internal processes for rapid dossier preparation. For small and medium-sized companies in particular, the resource burden and procedural uncertainty remain relevant concerns. The presentation will discuss what these lessons mean for BPI member companies and where industry, assessors and policymakers may identify overlapping problems and pragmatic solutions. Particular attention will be given to the interfaces between EU-level clinical assessment and national decision-making, the importance of predictable methodological guidance, and the need to ensure that European harmonisation translates into real procedural efficiency without creating additional layers of complexity.

13:30–15:00

Analysis and Modelling of Infectious Diseases in the Age of AI: Lessons from COVID-19

Mon, 28.09.26, 13:30–15:00, Room: K1

Chair: Martin Wolkewitz

Authors: Martin Wolkewitz [1]; Elizabeth Williamson [2]; Liesbeth de Wreede [3]; Hein Putter [3]

Affiliations: [1] Institut für Med. Biometrie und Med. Informatik, Freiburg; [2] London School of Hygiene and Tropical Medicine, London; [3] Leiden University Medical Center, Leiden

General outline and objectives

This workshop brings together leading experts to address the analysis and modelling of infectious diseases in the age of artificial intelligence, using the COVID-19 pandemic as a motivating example. The objective is to integrate perspectives from epidemiology, epidemic dynamic modelling, survival analysis, machine learning, and related methodological disciplines.

Rather than focusing on technical implementation or software details, the workshop emphasizes conceptual understanding, methodological challenges, interpretation, and data infrastructure requirements. Through concrete examples from the COVID-19 pandemic, the speakers will reflect on key analytical challenges, lessons learned, and future directions for infectious disease research in increasingly data-rich and AI-supported environments.

Structure

13:30 - 13:45 Johannes Bracher: Multi-model nowcasting and short-term forecasting of infectious diseases

13:45 - 14:00 Elizabeth Williamson: Identifying people at high-risk of severe COVID-19 outcomes

14:00 - 14:15 Simon Wood: The UK response to Covid-19: politics, models, goodthink, oldthink and statistics

14:15 - 14:30 Hein Putter: Estimating transmission parameters from stochastic epidemic models using survival analysis techniques

14:30 - 14:45 Liesbeth de Wreede: The impact of infection and vaccination on excess mortality analyzed by means of an extended multi-state model

14:45 - 15:00 Martin Wolkewitz / Tasnim Mafiz: Target trial emulation in infectious disease research: pitfalls and challenges

---

Multi-model nowcasting and short-term forecasting of infectious diseases

13:30-13:45

Presenter/Author: Johannes Bracher

Predictive epidemic modelling subsumes various tasks including nowcasting (the correction of reporting delays), short-term forecasting (unconditional statements about likely trajectories over short time horizons) and scenario modelling (longer-term projections under various sets of assumptions). In this talk, I will give an overview of the two first tasks, which play an important role for situational awareness. I will notably talk about multi-model approaches, where several models are run in parallel in order to assess agreement and generate combined ensemble forecasts. During the COVID-19 pandemic, these were organized in so-called "Forecast Hubs", which enabled the systematic evaluation of forecast performance. I will touch upon the statistical tools used to evaluate probabilistic forecasts and discuss challenges encountered in different types of prediction tasks. To conclude, I will briefly discuss the adaptation of these efforts as COVID-19 has ceased to be an emerging disease and become part of the broader landscape of respiratory infections.

Identifying people at high-risk of severe COVID-19 outcomes

13:45-14:00

Presenter/Author: Elizabeth Williamson

Affiliations: London School of Hygiene and Tropical Medicine, London

In the early days of the Covid-19 pandemic, one of the urgent questions for policy makers related to the identification of people at highest risk of severe outcomes from Covid-19 (hospitalization and death). Large scale routinely collected datasets, such as electronic health record data, provided a powerful source of information on risks experienced among the general population. The OpenSAFELY collaborative, an open source trusted research environment for electronic health record analysis, was set up in the UK in the early days of the Covid-19 pandemic. Linked data on Covid-19 tests, severe outcomes (hospitalization and death) and primary care electronic health records were made available for analysis within a secure environment. A fundamental feature of OpenSAFELY is that researchers never see the underlying data but send fully developed code into the secure live data environment to be executed against the real data, receiving back only summary results checked for disclosivity before release. Within the first few months, analyses were conducted through OpenSAFELY describing clinical and demographic characteristics of people who had experienced Covid-19 death, compared to the general population. This talk will present these early descriptive analyses designed to inform initial risk stratification, describe the OpenSAFELY workflow designed to ensure privacy, transparency and open working, and outline key challenges encountered in using these data. One way in which these early descriptive results, and similar results from other groups, were used by policy makers was to inform the selection of comorbidities which were placed on lists of high-risk individuals, often by choosing those resulting in large hazard ratios in multivariable models. This talk will discuss the challenge of informing risk stratification rules in a setting where a risk prediction model cannot be transported for use in the target population. Finally, this talk will explore the meaning of the question of "who is at highest risk?" and discuss how this question goes beyond the merely statistical and involves consideration of what is thought to be an "unfair" disparity in risk.

The UK response to Covid-19: politics, models, goodthink, oldthink and statistics

14:00-14:15

Presenter/Author: Simon Wood

The UK's Covid vaccine program, its properly conducted statistical Covid surveillance survey and its government's decision not to lock down in the face of Omicron (despite advise from modellers) were all world leading. But other aspects of the response were troubling from a statistical perspective, suggesting that risk management was far from optimal in the way that the Covid risks were balanced against the longer term risks from the response measures adopted. Once the unprecedented decision was taken to lock down the whole country, a dynamic appeared to develop in which the only politically acceptable opinion was that this decision had been both correct and unavoidable, with counter evidence ignored or even suppressed and very substantial effort devoted to supporting the acceptable narrative. This impeded revision of the policy approach. An especially unhelpful tendency was to view the predictions of epidemic models, never validated for prediction, not as mathematical statements of theories, but as scientific evidence somewhat superior to more empirically based evidence. This talk will review some of these issues, which are covered more fully in a Royal Statistical Society discussion paper (Wood et al. 2026, JRSSA 189: 1-138).

Estimating transmission parameters from stochastic epidemic models using survival analysis techniques

14:15-14:30

Presenter: Hein Putter

Authors: Hein Putter [1]; Jacco Wallinga [2]

Affiliations: [1] Leiden University Medical Center, Leiden; [2] National Institute for Public Health and the Environment, Bilthoven

Compartmental models based on ordinary differential equations (ODE's) quantifying the interactions between susceptible, infectious, and recovered individuals within a population have played an important role in infectious disease modeling. The aim of this presentation is to explain the link between stochastic epidemic models based on the susceptible-infectious-recovered (SIR) model, and methods from survival analysis. We illustrate how standard software for survival analysis in the statistical language R can be used to estimate pivotal parameters in the stochastic SIR model in the very much idealized situation where the epidemic is completely observed. Extensions incorporating interventions, age structure and heterogeneity are explored.

The impact of infection and vaccination on excess mortality analyzed by means of an extended multi-state model

14:30-14:45

Presenter: Eva Koster

Authors: Hein Putter; Liesbeth de Wreede

Affiliations: Leiden University Medical Center, Leiden

The COVID-19 pandemic has led to excess mortality worldwide. This 'excess mortality' is hard to quantify since it is different from the numbers of deaths due to COVID-19. Evaluating pandemic-related mortality should therefore not only be based on cause-of-death data but also on historical life tables to enable calculation of population-based measures of the difference between observed and expected mortality. We investigated the impact of infection with and vaccination against COVID-19 on excess mortality during 2020-2021 in persons aged 65 years or older in the Netherlands. We did this by incorporating relative survival modelling into a multi-state model with states vaccination, the acute and post-acute phase of (re)infection, and death, based on the recent methodology by Manevski et al. The multi-state model is a time-inhomogeneous Markov model, stratified by sex and age. In this model, transition probabilities can be estimated by the Aalen-Johansen estimator. Different from the standard relative survival setting, the excess hazard was a negative quantity for some groups during some periods, making regression modelling by means of a proportional hazards model infeasible. Modelling a pandemic made the choice of calendar time as timescale the most obvious, the consequences of which will be discussed, especially violation of the Markov property. We applied the model on real-world, nationwide and unselected individual data from Statistics Netherlands (CBS). From 1 January 2020 until the end of 2021, the total cumulative probability of excess mortality was 0.3%, 4.4% of all mortality. This percentage was markedly higher for older persons, especially men, both in absolute terms and as a percentage of subgroup-specific observed mortality. Almost all excess mortality took place after an infection but its probability was much lower for persons who had received a vaccination. The novel multi-state model incorporating relative survival enables to split all mortality in background and excess mortality with and without intermediate events. The current application shows the value outside the traditional context of relative survival for cancer patients.

Target trial emulation in infectious disease research: pitfalls and challenges

14:45-15:00

Presenter/Author: Martin Wolkewitz

Affiliations: Institut für Med. Biometrie und Med. Informatik, Freiburg

Target trial emulation has emerged as a major framework for estimating treatment effects from observational data in infectious disease research. During the early phase of the SARS-CoV-2 pandemic, it played a pivotal role in generating timely evidence when randomized controlled trials (RCTs) were

ethically, logistically, or practically infeasible. Beyond such emergency contexts, target trial emulation also serves broader objectives: (1) evaluating the external validity of RCT findings by assessing treatment effects in routine clinical practice, and (2) extending evidence to more heterogeneous populations and clinical settings, including rare outcomes and treatment switching. Drawing on findings from two reviews, we discuss key statistical challenges in target trial emulation for patients with COVID-19 and for patients with sepsis and/or hospital-acquired infections. We highlight common methodological pitfalls related to core design and analysis components, including the definition of time zero, the handling of time-related biases, competing events, and the specification of causal estimands. We further address challenges that are particularly relevant in infectious disease research, such as delays in initiating appropriate treatment, variation in treatment duration, and transitions from intravenous to oral therapy. In addition, we outline emerging methodological developments with potential relevance for infectious disease research. Approaches such as federated target trial emulation and the use of pre-trained neural networks for causal inference have shown promising performance in causal effect estimation and may outperform established methods in specific applications.

Between Data, Practice, and Evidence: (Inter)Nationally EHR Research

Mon, 28.09.26, 13:30–15:00, Room: K4

Chair: Veronika Strotbaum

Authors: Robin Grashof [1]; Saskia Kröner [2]; Veronika Strotbaum [3]; Björn Schreiweis [4]; Benjamin Kinast [4]

Affiliations: [1] Hochschule Niederrhein, Krefeld; [2] Hochschule Osnabrück - University of Applied Sciences, Osnabrück; [3] Working group "Consumer Health Informatics" within GMDS e.V., Berlin; [4] Institute for Medical Informatics and Artificial Intelligence, University Hospital Schleswig Holstein, Kiel

General outline and objectives

We invite participants to a collaborative workshop on national electronic health record (EHR) research. On one hand, we aim to assess the current status of (inter)national EHR Research across medical, technical, legal, governance, communicative and public health dimensions. In this context, we will identify research potential from identified gaps and needs. On the other hand, we aim to strengthen networking within German national EHR research by connecting the CHI working group with other GMDS-focus groups, researches outside the GMDS community and practitioners.

Structure

13:30 - 13:40 Robin Grashof, Saskia Kröner: A comparative Introduction to National Electronic Health Records: Germany's ePA and Turkey's e-Nabız

13:40 - 14:10 Benjamin Kinast, Björn Schreiweis: Group discussions concerning different national EHR research topics and relating research gaps

14:10 - 14:30

Group presentations and discussion in order to prioritize national EHR research

14:30 - 15:00

Plenary discussion to determinate next steps in German national EHR research

---

A comparative Introduction to National Electronic Health Records: Germany's ePA and Türkey's e-Nabız

13:30-13:40

Presenter/Author: Robin Grashof, Saskia Kröner

Affiliations: Hochschule Niederrhein University of Applied Sciences, Krefeld; Hochschule Osnabrück, Osnabrück

German national electronic health record (ePA) ("ePA für alle") was initiated in 2021 in Germany as a regulatory framework with a opt-in procedure that provided fundamental digital health services like

medical reports, manage practitioner access by citizens and medication management, to citizens who have explicitly consented to the activation of such services. The Consumer Health Informatics Working Group (AG CHI) of GMDS conducted a survey to ascertain the perspectives of citizens and healthcare professionals with regard to the benefits and barriers of ePA. The study found that consumers see the main benefits of the German ePA in collecting medical documents, improving data exchange between healthcare institutions, and providing emergency medical information, with high expectations for security, accessibility, and user-friendliness. However, adoption was limited by major barriers, especially insufficient information, lack of access, privacy concerns, perceived lack of necessity, time constraints, and limited technical understanding. Overall, the findings suggest that low ePA acceptance is strongly linked to a communication gap, and that direct information from health insurers and physicians, supported by multi-channel campaigns, may improve uptake. Following the introduction of the opt-out procedure in 2025, an individual ePA is established unless there is active objection. On the other side, Turkish national EHR "e-Nabız", introduced in 2015, is connected to an account for public services. e-Nabız is available via browser and smartphone application and comprehensively stores (text- and image-based) personal health data. The application provides multiple functions like data exchange between patients and multiple practitioners and long-term storage, point out individual health risks and provide related health information, manage practitioner access by citizens, making appointments, rating medical services, and getting requests for blood and bone marrow donations. National EHR research found that individual privacy concerns and perceived benefits regarding e-Nabız use are relevant for health data disclosure. In comparison between Turkish and German national EHR, both provide medical report exchange, privacy autonomy for citizens and medication management, whereas e-Nabız is more mature and empowers citizens to manage their health actively. While national EHRs vary from state to state, research at the national and international levels is needed to ensure high-quality, user-centered development.

Mapping metadata schemas to the HealthDCAT-AP standard: Lessons learned and hands on community examples (1/2)

Mon, 28.09.26, 13:30–15:00, Room: N3

Chair: Dagmar Waltemath, Judith Wodke, Sabine Oesterle

Authors: Judith Wodke [1]; Deepak Unni [2]; Vasundra Touré [2]; Matthias Löbe [3]; Núria Queralt Rosinach [4]; Sophie Klopfenstein [5]; Hanneke Leegwater [6]; Carina Nina Vorisek [6]; Sylvia Thun [6]; Thomas Ganslandt [7]; Dagmar Waltemath [8]; Sabine Oesterle [2]

Affiliations: [1] University Medicine Greifswald, Greifswald; [2] Swiss Institute of Bioinformatics, Basel; [3] Universität Leipzig, Leipzig; [4] Leiden University Medical Center, Leiden; [5] Berlin Institute of Health at Charité, Berlin; [6] Berlin Institute of Health @ Charité, Berlin; [7] Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen; [8] Universitätsmedizin Greifswald, Greifswald

General outline and objectives

According to the principles of Findable, Accessible, Interoperable, and Reusable (FAIR) data [1], research data should be made accessible accompanied by rich metadata that allows the assessment of their utility and reusability for a specific purpose beyond the original intent. With the advancement of the European Health Data Space (EHDS), European countries are required to harmonise their health data collections towards defined interoperability standards. The specified standard for metadata within the EHDS is the Health Data Catalog Vocabulary Application Profile (HealthDCAT-AP) [2].

While data formats and metadata schema differ between European countries and even between national initiatives, the challenge to harmonise metadata across countries is shared. The Swiss Personalized Health Network (SPHN), the Dutch Health-RI Initiative, the German National Research Data Infrastructure for Health (NFDI4Health) and Network University Medicine (NUM) / Medical Informatics Initiative (MII), and the FAIR Data Point specification used in several European funded projects are conducting a joint initiative to map their national metadata schemas to the HealthDCAT-AP standard.

In this workshop, we aim to present different approaches and share lessons learned. We encourage the communities of ISCB and GMDS to bring their own data examples and to work in a hands-on-hackathon on EHDS-compatible metadata enrichment for those data or our provided examples. We will engage in a lively discussion, sharing cross-border experiences and gathering feedback for further improvements of the provided mappings. We expect to align across biomedical disciplines the efforts towards interoperable metadata within the EU.

[1] Wilkinson MD, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar3(1):160018. doi: 10.1038/sdata.2016.18

[2] European Commission. HealthDCAT-AP Release 6 [Internet]. European Union 2025 [cited 2026 Feb 9]. Available from: https://healthdataeu.pages.code.europa.eu/healthdcat-ap/releases/release-6/

Structure

13:30 - 13:50 Workshop Introduction (moderation: Sabine Österle): Lightning talks from different European health data initiatives:

13:50 - 13:57 Sabine Österle: The SPHN Metadata Catalog: Reusing DCAT and HealthDCAT-AP to promote FAIR health metadata in Switzerland

13:57 - 14:04 Sophie A.I. Klopfenstein: Increasing the FAIRness of Studies and Related Resources across Health Research Domains - The NFDI4Health Metadata Schema

14:04 - 14:11 Thomas Ganslandt: Metadata in the German Medical Informatics Initiative & Network of University Medicine

14:11 - 14:18 Núria Queralt-Rosinach: FAIR Data Point: a strategic machine-actionable metadata service for autonomous discovery in the EHDS

14:18-14:25 Matthias Löbe: HealthDCAT-AP - Scope and usage scenarios in the European Health Data Space (EHDS)

14:25 - 14:45

Introduction of example data and group organisation (moderation: Deepak Unni, Judith Wodke)

14:45 - 15:30

The SPHN Metadata Catalog: Reusing DCAT and HealthDCAT-AP to promote FAIR health metadata in Switzerland

13:50-13:57

Presenter/Author: Deepak Unni, Vasundra Touré, Sabine Oesterle

Affiliations: SIB Swiss Institute of Bioinformatics, Basel

The growing availability of health datasets in Switzerland calls for dedicated frameworks supporting data discoverability and interoperability. This talk presents the Swiss Personalized Health Network (SPHN) Metadata Catalog, a solution based on the FAIR Data Point to enabling FAIR (findable, accessible, interoperable and reusable) management of metadata of health realted datasets across Switzerland. Built around the SPHN Metadata Catalog Schema, which reuses the Data Catalog Vocabulary (DCAT) and its health-domain extension HealthDCAT Application Profile (AP), the catalog ensures alignments with international standards while addressing specific clinical and research requirements. The SPHN Metadata Catalog offers both a human-facing interface for metadata browsing and a SPARQL endpoint for programmatic querying, enabling federation with other standard-compliant catalogs. We demonstrate how the SPHN Metadata Catalog contributes to a more connected and interoperable health data ecosystem in Switzerland and beyond, for international health data initiatives.

Increasing the FAIRness of Studies and Related Resources across Health Research Domains - The NFDI4Health Metadata Schema

13:57-14:04

Presenter/Author: Sophie Klopfenstein

Affiliations: Berlin Institute of Health at Charité, Berlin

To foster the FAIRness (Findable, Accessible, Interoperable, Reusable) of studies across diverse health research domains and to enable collaboration between clinical, epidemiological, and public health communities, the German National Research Data Infrastructure for Personal Health Data (NFDI4Health) has developed a modular metadata schema (MDS). This schema standardizes metadata from health studies and their related resources. Serving as the foundation for both the central Health Study Hub and Local Data Hubs, the NFDI4Health MDS supports generic metadata based on DataCite and domain-specific content based a. o. on ClinicalTrials.gov and the German Clinical Trials Register (DRKS). In light of the European Health Data Space entering into force, health data infrastructures face the critical challenge of ensuring compatibility with emerging European standards, specifically HealthDCAT-AP. We will explore strategic approaches to aligning the NFDI4Health MDS with these shifting regulatory and technical requirements, and discuss opportunities that this alignment creates for international health data interoperability and cross-border research collaborations.

Metadata in the German Medical Informatics Initiative & Network of University Medicine

14:04-14:11

Presenter/Author: Thomas Ganslandt

Affiliations: Friedrich-Alexander-Unviersität Erlangen-Nürnberg, Erlangen

In this talk, a short overview of metadata currently available from the Data Integration Centers of the German Medical Informatics Initiative and the Network of University Medicine will be given.

FAIR Data Point: a strategic machine-actionable metadata service for autonomous discovery in the EHDS

14:11-14:18

Presenter/Author: Núria Queralt Rosinach

Affiliations: Leiden University Medical Center, Leiden

The European Health Data Space (EHDS) mandates strict data sovereignty, meaning sensitive biomedical datasets must remain localized while remaining globally discoverable. This talk presents the FAIR

Data Point (FDP) as the strategic FAIR standard for this federated paradigm. By providing a common method for machine-actionable metadata, the FDP allows digital objects to be discovered and accessed autonomously by "FAIR-aware" agents. We examine the specific DCAT-based semantically-rich metadata and how it allows cross-institutional and cross-border query federation. For computational biologists, deploying FDPs is strategic since it bridges local data governance with the scale of a unified European research ecosystem, where digital objects such as datasets can be Findable, Accessible, Interoperable and Reusable (FAIR) for both humans and AI.

HealthDCAT-AP - Scope and usage scenarios in the European Health Data Space (EHDS)

14:18-14:25

Presenter/Author: Matthias Löbe

Affiliations: Universität Leipzig, Leipzig

HealthDCAT-AP is another domain-specific extension of the DCAT application profile for data portals in Europe. It aligns with the European Commission's general methodology for establishing European data spaces and ensures comprehensibility and interoperability. Annotating datasets with metadata according to HealthDCAT-AP is necessary but not automatically sufficient for easy discoverability and assessment of the data's fitness for purpose. A deeper understanding of the modeling constructs and alignment of best practices are required to unlock the full potential.

A Guided Experiment on Clinical Trial Data Sharing Quality and Data Reuse (1/2)

Mon, 28.09.26, 13:30–15:00, Room: N5

Chair: Ulrich Mansmann

Authors: Ulrich Mansmann [1]; Ka Hin Tai [2]; Giulia Varvarà [2]; Marcel Müller [3]; Nikolaus von Bomhard [4]

Affiliations: [1] Universität München, München; [2] University of Rennes, Inserm, EHESP, Irset - UMR_S 1085, Rennes; [3] Institut für Medizinische Informationsverarbeitung Biometrie und Epidemiologie (IBE), München; [4] Department of Medical Information Processing, Biometry, and Epidemiology, University of Munich, München

General outline and objectives

Clinical trial data sharing is increasingly mandated. Meaningful data reuse often fails due to inadequate data preparation, missing documentation, or misalignment between sharing modes and reuse needs. This workshop explores clinical trial data sharing and reuse and their technical implementation into a TRE (Trusted research environment). Besides practicing specific tools, the participants will anticipate challenges, trade-offs, and risks across the data sharing pipeline. Participants will develop a deeper understanding of how data quality and structure determine downstream reuse feasibility, FAIRness, and scientific value.

Structure

13:30 - 14:15 Ka-Hin Tai, Ulrich Mansmann: A Sharing Intention Is Not Enough: Guided Experiments on Clinical Trial Data Quality and Reuse - Data sharing

14:15 - 15:00 Giulia Varvara: A Sharing Intention Is Not Enough: Guided Experiments on Clinical Trial Data Quality and Reuse - Data reuse

---

A Sharing Intention Is Not Enough: Guided Experiments on Clinical Trial Data Quality and Reuse - Data sharing

13:30-14:15

Presenter: Ulrich Mansmann

Authors: Ulrich Mansmann [1]; Ka Hin Tai [2]

Affiliations: [1] Universität München, München; [2] Université de Rennes, Rennes

Introduction: Clinical trial data sharing is widely promoted, yet the practical effort required to make data meaningfully shareable is often underestimated. This session presents a guided thought experiment that places participants in the role of data generators, confronting the operational realities of preparing clinical trial data for sharing. Methods: Using fictional and real clinical trial data

(https://pubmed.ncbi.nlm.nih.gov/ 37414064/), participants are exposed to different representations of the same study, ranging from unstructured records to partially documented datasets and more structured, FAIR-aligned formats. They are asked to assess shareability under different dissemination modes, such as direct sharing on request, repository deposition, or access through controlled environments. Rather than focusing on ideal standards, the exercise emphasizes decision-making under constraints: what preparation work is required, what cannot reasonably be shared, and which governance or ethical risks must be addressed. Results: The thought experiment highlights key tensions between effort, usability, and control. Participants explore how documentation quality, metadata completeness, and data structure influence the feasibility of sharing, and where diminishing returns may arise in striving for full FAIR compliance. The discussion also surfaces common blind spots, including missing context, unclear variable definitions, and the implicit assumptions embedded in data collection processes. Conclusions: By comparing different scenarios, this session demonstrates that "shareable" data is not a binary state but a spectrum shaped by intended use, sharing mode, and available resources. The central insight is that data sharing must be designed with downstream use in mind, rather than treated as a post hoc obligation. This guided exercise equips participants with a practical framework to anticipate preparation effort, make defensible trade-offs, and develop proportionate data sharing strategies in real-world clinical research settings. The session is aimed at clinical researchers, biostatisticians, data stewards, and anyone involved in preparing or managing clinical trial research data. Required reading includes 10.1016/j.jclinepi.2025.111920.

A Sharing Intention Is Not Enough: Guided Experiments on Clinical Trial Data Quality and Reuse - Data reuse

14:15-15:00

Presenter: Ulrich Mansmann, Giulia Varvarà

Authors: Giulia Varvarà

Affiliations: University of Rennes, Inserm, EHESP, Irset - UMR_S 1085, Rennes

The success of clinical trial data sharing is ultimately determined by whether shared data can be meaningfully reused. This session serves as a sequel to our first session and presents a guided thought experiment that shifts perspective from data generators to data reusers, focusing on the practical challenges of working with shared clinical trial data. We orientate the discussion around the Lancet depression trial (https://pubmed.ncbi.nlm.nih.gov/37414064/). Participants are assigned realistic reuse scenarios, such as re-analysis, secondary analysis, or inclusion in meta-analysis, and are asked to work with conceptually "shared" data packages prepared under varying assumptions. The task is to identify the minimum viable data required for their analytical goal, to draft a data request application, assess the usability of the available materials, and determine what additional information or clarification is needed. Through this exercise, participants encounter common barriers to reuse, including incomplete metadata, ambiguous variable definitions, missing documentation, and uncertainty around data provenance. They also reflect on how different reuse purposes impose distinct requirements on data structure, granularity, and contextual information. By comparing across

scenarios, participants observe that data suitable for one type of analysis may be inadequate for another. The discussion emphasizes that reuse feasibility is shaped by upstream decisions made during data preparation and sharing. Small omissions in documentation or structure can propagate into substantial inefficiencies or even render data unusable. Conversely, well-designed data packages can significantly lower the barrier to reuse and expand the range of possible analyses. This session demonstrates that data reuse is on how to get access and how to align data data received with analytical intent. By engaging in structured reasoning, participants develop a clearer understanding of how to apply for data, evaluate shared data, anticipate reuse challenges, and advocate for reuse-aware data sharing practices that maximize scientific value. The session is aimed at researchers, methodologists, and analysts interested in evaluating and reusing shared clinical data. Required reading includes 10.1016/j.jclinepi.2025.111984.

15:30–17:00

Uncertainty in Clinical Predictions

Mon, 28.09.26, 15:30–17:00, Room: K1

Chair: Ewout Steyerberg

Authors: Ewout Steyerberg [1]; Doranne Thomassen [2]; Ben Van Calster [3]

Affiliations: [1] Julius Center, Utrecht; [2] Biomedical Data Sciences, LUMC, Leiden; [3] KU Leuven, Leuven

General outline and objectives

Clinical prediction models for a health condition are commonly evaluated regarding performance or utility for a population, although decisions are made for individuals. In this workshop, we first discuss developments in quantification of finite-sample uncertainty for individual patients in terms of 'effective sample size': the number of similar patients that a prediction is effectively based on. This number of 'patients-like-you' can be estimated for predictions based on a wide range of models, including generalized linear models and various machine learning methods. Novel developments include the estimation of effective sample size for treatment benefit at the individual level. We discuss implications for model development and transparent risk communication.

Uncertainty in risk estimates for individuals also relates to model uncertainty (variability in modeling choices made by different modelers) and applicability uncertainty (variability in measurement procedures and between populations). We will present case studies and simulations to illustrate the huge role of model uncertainty and applicability uncertainty.

Structure

15:30 - 15:45 Ewout Steyerberg: Introduction and Overview on Uncertainty in predictions: from quantification to sample size challenges

15:45 - 16:15 Doranne Thomassen: Quantifying finite-sample uncertainty by effective sample size for predictions

16:15 - 16:45 Ben van Calster: Beyond sample size: model and applicability uncertainty

16:45 - 17:00

Interactive Panel Discussion: Moderated discussion with all presenters addressing audience questions, conceptual and implementation challenges, and future directions including collaborative opportunities.

Quantifying finite-sample uncertainty in predictions using the effective sample size

15:45-16:15

Presenter/Author: Doranne Thomassen

Affiliations: Leiden University Medical Center, Leiden

Uncertainty around clinical risk predictions should be evaluated and made transparent, as a prediction algorithm may be sufficiently trustworthy for some patients but not others. We will discuss the effective sample size as a measure of sampling uncertainty in individual risk predictions based on regression and machine learning models, with recent extension to causal prediction models. We show how effective sample size can be obtained in various prediction settings and discuss illustrations of the (causal) effective sample size in two clinical datasets. For one dataset (n=23,034) with patients suffering from acute myocardial infarction, we aimed to predict 30-day mortality using five types of risk prediction models: logistic regression, elastic net, XGBoost, neural net, and random forest. While traditional performance metrics such as the concordance statistic were highly similar across models, there were large differences in effective sample sizes and risk predictions at the individual level. In a dataset of primary care patients with type 2 diabetes (n=15,743), we developed causal prediction models for outcomes under second-line treatment options. Here, confounding adjustment variables affected the effective sample size in addition to the predictor variables. Between-model differences in the distribution of effective sample sizes pointed to tradeoffs between assumptions of no unmeasured confounding and correct model specification and positivity. As in the MI dataset, large between-individual differences in effective sample size were observed. In conclusion, prediction uncertainty at the individual level can be substantial even when predictions algorithms are developed with large sample size. The effective sample size contributes to an individual-level evaluation of prediction uncertainty, improving the transparency of (causal) predictions. References: Thomassen D, le Cessie S, van Houwelingen HC, Steyerberg EW. Effective sample size: A measure of individual uncertainty in predictions. Statistics in Medicine. 2024 1-13. Thomassen D, Hackmann T, Goeman JJ, Steyerberg EW, le Cessie S. Effective sample size for individual risk predictions: quantifying uncertainty in machine learning models. The Lancet Digital Health, 2025 7 Van Geloven N, Thomassen D, Weir DL, Kunneman M. Does this algorithm apply to me? Developing causal effective sample size for decision support algorithms - project funded by Dutch Research Council (NWO) under grant DOI 10.61686/DFECP93059.

Beyond sample size: model and applicability uncertainty

16:15-16:45

Presenter: Ben Van Calster

Authors: Ben Van Calster; Lasai Barreñada

Affiliations: KU Leuven, Leuven

Clinical prediction models for a health condition are commonly evaluated regarding performance for a population, although decisions are made for individuals. The classic view relates uncertainty in risk estimates for individuals to sample size (estimation uncertainty) but uncertainty can also be

caused by model uncertainty (variability in modeling choices) and applicability uncertainty (variability in measurement procedures and between populations). We used real and synthetic data for ovarian cancer diagnosis to train 59,400 models with variations in estimation, model, and applicability uncertainty. We estimate the probability of ovarian cancer for a fixed test set of 100 patients according to each model. We find that estimation uncertainty can be dominated by model uncertainty and applicability uncertainty, even for models that perform well at the population level. Estimation uncertainty decreased considerably with increasing training sample size, whereas model and applicability uncertainty remained large. In conclusion, individual risk estimates are more uncertain than often assumed due to model uncertainty and applicability uncertainty. Predictive algorithms should inform, not dictate, care and support personalization through clinician-patient interaction rather than through inherently uncertain model outputs. References Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. Journal of the Royal Statistical Society Series A: Statistics in Society, 158(3), 419-444. Riley, R. D., Collins, G. S., Kirton, L., Snell, K. I., Ensor, J., Whittle, R., ... & Archer, L. (2025). Uncertainty of risk estimates from clinical prediction models: rationale, challenges, and approaches. BMJ, 388. Barreñada, L., Steyerberg, E. W., Timmerman, D., Thomassen, D., Wynants, L., & Van Calster, B. (2025). The fundamental problem of risk prediction for individuals: health AI, uncertainty, and personalized medicine. arXiv preprint arXiv:2506.17141. This research is supported by Research Foundation Flanders (FWO) grants G097322N, G049312N, G0B4716N, and 12F3114N KU Leuven internal grants C24M/20/064 and C24/15/037

Large Language Model Informed Real-World Cohort Data

Mon, 28.09.26, 15:30–17:00, Room: K4

Chair: Nadine Binder

Authors: Michelle Pfaffenlehner [1]; Nadine Binder [2]; Harald Binder [1]; Prasenjit Mitra [3]; James Carpenter [4]

Affiliations: [1] University of Freiburg, Freiburg, Germany, Freiburg; [2] Medical Center - University of Freiburg, Freiburg im Breisgau; [3] Carnegie Mellon University Africa, Kigali; [4] London School of Hygiene & Tropical Medicine, London

General outline and objectives

Using large language models (LLMs) in the context of real-world cohort data modeling or analysis offers new opportunities and a wide range of applications. This workshop provides insights into the integration of LLMs with real-world data (RWD) through three illustrative use cases.

First, we present the use of LLMs to generate synthetic data, introducing a structured framework that defines key principles and steps for designing such datasets. Second, we discuss the interaction of LLMs with synthetic data, focusing on data enrichment and assessing how LLMs interpret and respond to real-world data structures. Third, we examine the application of LLMs for the analysis of RWD, highlighting potential applications, limitations and methodological considerations.

Structure

15:30 - 15:50 Harald Binder: Leveraging large language models and agents for synthetic longitudinal disease cohort data in an expert-driven approach

15:50 - 16:10 Michelle Pfaffenlehner: Evaluating LLM-predictions of exacerbations in chronic obstructive pulmonary disease with a synthetic benchmark from routine clinical data

16:10 - 16:30 Prasenjit Mitra: Errors by LLMs: Do we need a reporting system?

16:30 - 17:00

Fishbowl Panel Discussion with speakers and panelist James Carpenter

---

Leveraging large language models and agents for synthetic longitudinal disease cohort data in an expert-driven approach

15:30-15:50

Presenter/Author: Harald Binder

Affiliations: University of Freiburg, Freiburg, Germany, Freiburg

Synthetic data strategies in clinical cohort settings typically rely on real data and/or expert knowledge for specifying underlying data structure, but gathering appropriate real data often is difficult,

and expert knowledge difficult to formalize. Large language models (LLMs), which incorporate a huge body of knowledge from the biomedical literature, might be useful for addressing both challenges as a kind of synthetic expert. However, it is unclear so far how to best extract corresponding information, i.e. on data structure or biomedical knowledge, in a stable, reproducible manner. Agentic systems, which can combine LLMs with data generation and analysis tools in an iterative way, promise a way for implementing evaluation criteria, but also render the setup more complex. I will illustrate different ways for extracting information in longitudinal disease cohort data settings from LLMs, with a focus on pair and triplet questions, and using agentic systems, where information extraction is phrased as a programming challenge to the agent. As a unifying theme, I will discuss implicit assumptions with respect to problem and world representations underlying LLMs and agentic systems, and how these influence the quality of the resulting synthetic data.

Evaluating LLM-predictions of exacerbations in chronic obstructive pulmonary disease with a synthetic benchmark from routine clinical data

15:50-16:10

Presenter/Author: Michelle Pfaffenlehner

Affiliations: Institut für medizinische Biometrie und Statistik, Universität Freiburg, Freiburg

Large Language Models (LLMs) demonstrate potential in predicting clinical events from patient histories, yet robust evaluation methods for such tasks remain underdeveloped. To address this, we focus on predicting acute exacerbations in chronic obstructive pulmonary disease (COPD). We constructed a synthetic benchmark dataset using a Hidden Markov Model, providing a known ground truth of underlying patient states to rigorously evaluate predictive performance. Using a pair-wise comparison task, we evaluate whether LLMs can identify which of two patients is more likely to be readmitted to the hospital with a COPD exacerbation. Our experimental design compares local models, i.e., gpt-oss-120b and gemma-4-26b-a4b-it, and systematically tests various prompting strategies. We examine a wide array of prompt elements, including current and previous diagnoses (ICD-10-GM), procedure codes (OPS), and length of stay (LOS), to determine which clinical information most improves predictive accuracy. By analyzing the interaction between these prompt elements through logistic regression, we identify the most significant predictors of model success. This approach allows us to uncover critical performance differences between prompting strategies and model architectures. More generally, this work provides a foundational framework for benchmarking LLMs in clinical settings and motivates further research into data-driven prediction of COPD progression.

Errors by LLMs: Do we need a reporting system?

16:10-16:30

Presenter/Author: Prasenjit Mitra

Affiliations: Carnegie Mellon University Africa, Kigali

In this talk we will talk about the types of errors a large language model frequently makes. The ways in which they fail are poorly understood and active areas of investigation: hallucinations, reasoning collapse, mis-calibrated confidence, and input brittleness to name a few Research in these areas have been pursued in isolation, producing fragmented mitigation strategies. We argue that this fragmentation is unnecessary and call for more integrated investigations. We surveyed work on this topic and collected them in a unified uncertainty-driven diagnostic matrix that maps observable LLM failure modes to a set of shared internal uncertainty primitives revealing deep structural commonalities across seemingly unrelated error categories. Building on this mapping, we call for a root-cause identification framework that shifts the focus from symptom-level fixes to interventions targeting upstream uncertainty sources. This perspective shows that a single targeted intervention can systematically resolve entire clusters of downstream failures. We hope to provide medical practitioners to detect such failures and mitigate them using a principled basis using uncertainty root cause rather than failure symptom, and provide researchers with a unified diagnostic vocabulary for building more reliable LLM systems. Unfortunately, we do not know the harm LLMs' medical advice causes to lay users. Drugs, vaccines, and devices have mature safety surveillance systems - exemplified by FAERS and VAERS, now consolidated into the FDA's Adverse Event Monitoring System (AEMS) - which turn individual reports into population-level signals and have saved countless lives. General-purpose conversational AI operates outside this infrastructure. No one is counting. In the second part of this talk, we propose an adverse event reporting system designed for this gap. It comprises a taxonomy that classifies realized harms and near misses across clinical, behavioral, emotional, and equity dimensions a tiered privacy model that lets reporters choose their level of identifiability and a data architecture that routes reports into structured stores from which population-level safety signals can be detected and distinguished from early warnings. The system does not by itself close the blind spot, but it makes it visible, and gives policymakers, regulators, and developers the evidence they currently lack.

Mapping metadata schemas to the HealthDCAT-AP standard: Lessons learned and hands on community examples (2/2)

Mon, 28.09.26, 15:30–17:00, Room: N3

Chair: Dagmar Waltemath, Judith Wodke

Authors: Judith Wodke [1]; Deepak Unni [2]; Vasundra Touré [2]; Matthias Löbe [3]; Núria Queralt Rosinach [4]; Sophie Klopfenstein [5]; Hanneke Leegwater [6]; Carina Nina Vorisek [6]; Sylvia Thun [6]; Thomas Ganslandt [7]; Dagmar Waltemath [8]; Sabine Oesterle [2]

Affiliations: [1] University Medicine Greifswald, Greifswald; [2] Swiss Institute of Bioinformatics, Basel; [3] Universität Leipzig, Leipzig; [4] Leiden University Medical Center, Leiden; [5] Berlin Institute of Health at Charité, Berlin; [6] Berlin Institute of Health @ Charité, Berlin; [7] Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen; [8] Universitätsmedizin Greifswald, Greifswald

General outline and objectives

According to the principles of Findable, Accessible, Interoperable, and Reusable (FAIR) data [1], research data should be made accessible accompanied by rich metadata that allows the assessment of their utility and reusability for a specific purpose beyond the original intent. With the advancement of the European Health Data Space (EHDS), European countries are required to harmonise their health data collections towards defined interoperability standards. The specified standard for metadata within the EHDS is the Health Data Catalog Vocabulary Application Profile (HealthDCAT-AP) [2].

While data formats and metadata schema differ between European countries and even between national initiatives, the challenge to harmonise metadata across countries is shared. The Swiss Personalized Health Network (SPHN), the Dutch Health-RI Initiative, the German National Research Data Infrastructure for Health (NFDI4Health) and Network University Medicine (NUM) / Medical Informatics Initiative (MII), and the FAIR Data Point specification used in several European funded projects are conducting a joint initiative to map their national metadata schemas to the HealthDCAT-AP standard.

In this workshop, we aim to present different approaches and share lessons learned. We encourage the communities of ISCB and GMDS to bring their own data examples and to work in a hands-on-hackathon on EHDS-compatible metadata enrichment for those data or our provided examples. We will engage in a lively discussion, sharing cross-border experiences and gathering feedback for further improvements of the provided mappings. We expect to align across biomedical disciplines the efforts towards interoperable metadata within the EU.

[1] Wilkinson MD, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar3(1):160018. doi: 10.1038/sdata.2016.18

[2] European Commission. HealthDCAT-AP Release 6 [Internet]. European Union 2025 [cited 2026 Feb 9]. Available from: https://healthdataeu.pages.code.europa.eu/healthdcat-ap/releases/release-6/

Structure

15:30 - 16:30

Hands-on-session: EHDS-compatible metadata enrichment for example data (moderation: Deepak Unni, Judith Wodke)

16:30 - 17:00

Discussion (moderation: Dagmar Waltemath)

A Guided Experiment on Clinical Trial Data Sharing Quality and Data Reuse (2/2)

Mon, 28.09.26, 15:30–17:00, Room: N5

Chair: Ulrich Mansmann

Authors: Ulrich Mansmann [1]; Ka Hin Tai [2]; Giulia Varvarà [2]; Marcel Müller [3]; Nikolaus von Bomhard [4]

Affiliations: [1] Universität München, München; [2] University of Rennes, Inserm, EHESP, Irset - UMR_S 1085, Rennes; [3] Institut für Medizinische Informationsverarbeitung Biometrie und Epidemiologie (IBE), München; [4] Department of Medical Information Processing, Biometry, and Epidemiology, University of Munich, München

General outline and objectives

Clinical trial data sharing is increasingly mandated. Meaningful data reuse often fails due to inadequate data preparation, missing documentation, or misalignment between sharing modes and reuse needs. This workshop explores clinical trial data sharing and reuse and their technical implementation into a TRE (Trusted research environment). Besides practicing specific tools, the participants will anticipate challenges, trade-offs, and risks across the data sharing pipeline. Participants will develop a deeper understanding of how data quality and structure determine downstream reuse feasibility, FAIRness, and scientific value.

Structure

15:30 - 16:30 Marcel Müller, Nikolaus von Bomhard: A Sharing Intention Is Not Enough: Guided Experiments on Clinical Trial Data Quality and Reuse - TRACE showcase

16:30 - 17:00 Moderation: Ulrich Mansmann: Discussion with speakers and participants

---

A Sharing Intention Is Not Enough: Guided Experiments on Clinical Trial Data Quality and Reuse - TRACE showcase

15:30-16:30

Presenter/Author: Marcel Müller, Nikolaus von Bomhard

Affiliations: Ludwig-Maximilians-Universität München Institute for Medical Information Processing, Biometry, and Epidemiology (IBE), Muncih; Ludwig-Maximilians-Universität München Institute for Medical Information Processing, Biometry, and Epidemiology (IBE), Munich

Introduction: Translating principles of clinical trial data sharing and reuse into practice requires infrastructures that integrate governance, data management, and analytical workflows. This showcase presents TRACE (Trusted Research Access & Collaboration Environment) as an implementation of a trusted research environment (TRE). Methods: The showcase uses original data from the Lancet Depression trial (https://pubmed.ncbi.nlm.nih.gov/37414064/) and demonstrates how TRACE supports

the end-to-end process of accessing, managing, and analysing clinical trial data within a controlled environment. Key features include structured data integration, metadata representation, governed access procedures, and embedded analytical workspaces. Participants are guided through how data packages of varying quality are handled within the system, from initial access requests to exploratory analysis and output generation. Results: The demonstration highlights how design choices within a TRE influence the usability and reproducibility of analyses. For example, the availability of data dictionaries, clarity of coding schemes, and completeness of metadata directly affect the efficiency with which users can understand and work with the data. The showcase also illustrates how governance mechanisms, such as access controls and audit trails, interact with analytical flexibility. By linking back to the presentations on data sharing and reuse, we provide insights how upstream data preparation decisions manifest in a real-world environment. Well-prepared datasets enable smooth navigation and rapid analysis, while poorly documented data introduce friction and uncertainty. Conclusion: The central message is that TREs do not resolve data quality issues but make them explicit. This showcase helps participants understand how technical systems, governance requirements, and data design jointly determine whether shared clinical trial data becomes a usable scientific resource. The session is aimed at researchers, data stewards, and informaticians interested in practical infrastructures for governed data access and analysis. No prior preparation is required.

Tue 29.09.2026

11:00–12:30

Reliable Subgroup Identification and Analysis - Results of GMDS Biostatistics Competition 2026

Tue, 29.09.26, 11:00–12:30, Room: K1

Chair: Anika Großhennig

Authors: Max Westphal [1]; Anika Großhennig [2]; Heiko Götte [3]; Aslihan Gerhold-Ay [3]; Clara Pavillet [3]

Affiliations: [1] Fraunhofer MEVIS, Bremen; [2] Medizinische Hochschule Hannover, Hannover; [3] Merck Healthcare KGaA, Darmstadt

General outline and objectives

The GMDS Biostatistics Competition 2026, organized by the Biometrics section of the GMDS in collaboration with Merck Healthcare KGaA, addresses the critical challenge of reliable subgroup identification in clinical trial data. With 30 simulated datasets that closely mirror real-world phase II/III randomized trials, the competition aims to evaluate methodological approaches for detecting treatment-effect heterogeneity, identifying predictive and prognostic biomarkers, assigning patients to subgroups, estimating subgroup proportions, and quantifying treatment effects within identified groups. Participants are challenged to developing transparent, reproducible, and robust analytical workflows that balance statistical power with interpretability, moving beyond black-box machine learning toward clinically meaningful and interpretable approaches. The data-generating mechanism and true underlying subgroup assignment remain undisclosed, ensuring a rigorous and unbiased evaluation. Performance is assessed using task-specific metrics, including agreement rates, variable selection accuracy, subgroup proportion estimation, and treatment effect estimation via RMSE, compared against the true data-generating model.

In this workshop, we present the overarching results of the competition, highlight the most effective and innovative approaches, and showcase exemplary solutions from selected teams that combine methodological rigor, reliability and practical relevance. The session will be streamed live, enabling global participation and fostering open dialogue on the future of subgroup analysis in evidence-based medicine.

Structure

11:00-11:15 Max Westphal: Introduction to the Biostatistics Competition

11:15-11:30 Heiko Götte, Aslihan Gerhold-Ay, Clara Eléonore Pavillet: Overall results and findings of the Biostatistics Competition

11:30-12:15

Presentations of solutions by participants of the competition

12:15-12:30

Award ceremony and discussion

Introduction to Individual Participant Data Meta-Analysis

Tue, 29.09.26, 11:00–12:30, Room: K4

Chair: Tim Mori, Verena S Hoffmann

Authors: Verena S Hoffmann [1]; Tim Mori [2]; Manuel Feißt [3]; Alexandra Strobel [4]; Christian Röver [5]; Tim Mathes [6]; Tim Friede [5]

Affiliations: [1] Ludwig-Maximilians-Universität München Institute for Medical Information Processing, Biometry, and Epidemiology (IBE), München; [2] Deutsches Diabetes-Zentrum (DDZ), Leibniz-Zentrum für Diabetes-Forschung an der Heinrich-Heine-Universität Düsseldorf, Düsseldorf; [3] Institute of Medical Biometry, Heidelberg, Heidelberg; [4] Martin-Luther-University Halle-Wittenberg, Halle-Wittenberg; [5] Universitätsmedizin Göttingen, Göttingen; [6] Institut für Medizinische Statistik, Universitätsmedizin Göttingen, Göttingen

General outline and objectives

Individual participant data (IPD) meta-analysis has become an essential tool in modern evidence synthesis, particularly in clinical research, regulatory decision-making, and health technology assessment. Compared with conventional aggregate data meta-analysis, IPD approaches enable more flexible modelling of intervention effects, improved handling of rare outcomes, and robust investigation of subgroup effects. The primary objective of this workshop is to provide participants with a clear conceptual and methodological understanding of IPD meta-analysis, its advantages, limitations, and appropriate use cases, thereby supporting rigorous and transparent evidence generation.

Structure

The workshop begins by motivating the use of IPD meta-analysis and critically discussing the limitations of naïve pooling approaches, particularly the risks of bias arising from unaccounted study-level differences. We then introduce one-stage hierarchical models that properly account for clustering of participants within studies and distinguish between within-trial and between-trial interactions and discuss strategies for exploring effect modification. Subsequently, two-stage IPD methods are covered, including meta-regression and more advanced strategies for investigating treatment effect modifiers. Throughout, methodological rigor and transparent modelling choices are emphasized.

On the Ethical and Technical Aspects for Deployment and Evaluation of AI-Driven Clinical Decision Support Systems

Tue, 29.09.26, 11:00–12:30, Room: N5

Chair: Stefan Rühlicke, Zully Ritter

Moderators: Zully Ritter and Stefan Rühlicke

Authors: Zully Ritter [1]; Stefan Rühlicke [2]; Sebastian Fudickar [3]; Maximillan Sprang [4, 5]; Srushhti Trivedi [6, 7]; Roland Roller [8]; Sobhan Moazemi [9]; Rosa E. Martín-Peña [10, 11]; Sheikh Faisal Rashid [12, 13]; Gabriela Botea [14]

Affiliations: [1] University Medical Center Göttingen, Göttingen; [2] Universitätsklinikum Jena, Jena; [3] Universität zu Lübeck, Lübeck; [4] Dept of Dermatology, University Medical Center Mainz, Mainz; [5] IQCB, Johannes Gutenberg University, Mainz; [6] Department of Medical Informatics, University Medical Center Göttingen (UMG), Göttingen; [7] Department of Medical Ethics and History in Medicine, University Medical Center Göttingen, Göttingen; [8] German Research Center for Artificial Intelligence (DFKI), Berlin, Germany, Berlin; [9] Helmholtz Centre for Infection Research, Braunschweig; [10] CELLS - Centre for Ethics and Law in the Life Sciences, Leibniz Universität Hannover, Hannover; [11] CAIMed - Center for AI and Causal Methods in Medicine, Hannover; [12] German Research Center for Artificial Intelligence DFKI, Berlin; [13] International University of Applied Sciences, Berlin; [14] Google Research, Berlin

General outline and objectives

Although AI techniques such as Machine Learning, NLP, and Computer Vision are widely used in healthcare for diagnosis, disease management, and risk prediction, far less attention has been paid to ethical metrics in their technical development. This workshop presents methods for measuring and evaluating data ethics, focusing on defining, quantifying, and distinguishing ethical biases and noise from biases and noise in data and algorithms.

Structure

11:00 - 11:05 Zully Ritter and Stefan Rühlicke: Welcome

11:05 - 11:13 Roland Roller: The Fairy Tale of Fair Clinical AI

11:13 - 11:21 Rosa-Esther Martin Peña: The Uncoded Patient: Structural Blind Spots in Clinical AI Evaluation

11:21 - 11:29 Sobhan Moazemi: Features Mutual Information (FMI) and eXplainable Information Criterion (XIC): A Concept for Novel Objective Metrics for eXplainable Artificial Intelligence in Medicine Leveraging Information Theory

11:29 - 11:37 Faisal Rashid: Bridging Ethical and Technical Evaluation in Machine Learning for Clinical Decision

Support Systems: A Data-Centric Perspective

11:37 - 11:45 Maximilian Sprang: Overlooked sample quality imbalance between patient groups impairs reproducibility of clinically relevant RNA-seq studies.

11:45 - 11:54 Srushhti Trivedi: Navigating the Ethical Landscape of Artificial Intelligence in Clinical Applications: The Role of Algorithmic Impact Assessment Tool

11:54 - 12:24 Moderators: Zully Ritter and Stefan Rühlicke: Panel Discussion

12:24 - 12:30

Closing engagement

The Fairy Tale of Fair Clinical AI

11:05-11:12

Presenter/Author: Roland Roller

Affiliations: Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Berlin

Machine learning and large language models (LLMs) are increasingly integrated into clinical decision support systems (CDSS), often demonstrating strong overall performance on benchmark datasets. However, aggregate evaluation metrics can conceal substantial performance disparities across patient subgroups, creating risks for vulnerable or underrepresented populations. In this talk, we argue that the idea of universally "fair" clinical AI systems is, at least currently, closer to a fairy tale than a realistic deployment assumption. Instead of assuming fairness from overall benchmark performance, we advocate for systematic subgroup-oriented evaluation and transparent reporting of model limitations before clinical deployment. Building on previous work, we present findings from multiple multimodal medical prediction tasks, including mortality prediction, triage support, and graft failure prediction. Across these tasks, we observe that model behavior varies considerably depending on patient characteristics such as age, sex, ethnicity, data completeness, and clinical context. Importantly, these disparities often remain hidden when only overall accuracy or AUC metrics are reported. The talk further expands the discussion beyond model architectures themselves and highlights how bias can emerge throughout the entire NLP and clinical AI pipeline. This includes data collection, annotation practices, modality imbalance, missing information, preprocessing decisions, and downstream evaluation choices. In particular, we discuss the often-overlooked fairness implications of anonymization and de-identification systems. While privacy-preserving methods are essential in medical AI, they may disproportionately affect underrepresented populations by removing rare linguistic patterns, demographic signals, or clinically relevant contextual information. As a result, anonymization pipelines themselves can unintentionally contribute to subgroup-specific performance degradation. Finally, we briefly reflect on the limitations of current bias mitigation approaches. Many methods assume that disparities can be reduced through post-hoc adjustments or fairness constraints on the model level. However, such approaches are often constrained by data imbalance, conflicting definitions of fairness, and trade-offs with predictive performance or privacy requirements. This highlights the need for a more holistic view: rather than treating bias mitigation as a final fix, it should be embedded in a broader framework of continuous evaluation, transparency, and context-specific deployment decisions in clinical practice.

The Uncoded Patient: Structural Blind Spots in Clinical AI Evaluation

11:12-11:19

Presenter/Author: Rosa E. Martín-Peña

Consider Leo: a worker with cancer whose treatment fatigue is reinterpreted by algorithmic dashboards as cognitive decline, leading to her dismissal. The system does not malfunction, it functions perfectly within its ontology. It can only see what has been coded as signal. Her illness, the lived experience of treatment, and her sickness, the cultural decision not to disclose, produce what this contribution calls double noise: the convergence of algorithmic distortion and epistemic exclusion in the absence of feedback. Nowconsider a different system, a decade earlier: the PACE trial enrolling ME/CFS patients in graded exercise therapy. That trial, too, functioned perfectly within its ontology. It could not represent post-exertional malaise not because the data were absent, but because the mechanism had no node in the causal graph. Patients deteriorated. The harm was not an error within the framework. It was the predictable consequence of a framework never adequate for the condition it claimed to govern, and that framework is now inherited by default by every clinical decision support system trained on the data it produced. These two cases instantiate the same structural failure. Drawing on the classical sociological triad of disease, illness, and sickness, this contribution argues that CDSS systematically exclude illness and sickness not through oversight but through design: clinical documentation, ICD coding infrastructures, and informed consent frameworks are built around disease as a measurable biological entity. Illness and sickness are either absent or petrified into formats that strip their meaning. Crucially, informed consent, designed to protect the patient, filters out precisely the experiential data that would make CDSS adequate. Current ethical benchmarks cannot detect these exclusions because they operate within the same ontological substrate that produces them. This contribution presents a typology of epistemic exclusions organized around the disease/illness/sickness triad and invites discussion on what targeted corrective interventions, regulatory, institutional, and technical, would need to address them.

Features Mutual Information (FMI) and eXplainable Information Criterion (XIC): A Concept for Novel Objective Metrics for eXplainable Artificial Intelligence in Medicine Leveraging Information Theory

11:19-11:26

Presenter/Author: Sobhan Moazemi

Affiliations: Hemlholtz Centre for Infection Research, Braunschweig

Artificial Intelligence (AI) has faced impressive advancements in recent years resulting in paradigm shifts in many application domains such as Medicine. However, the inherent high-stakes nature of decision making in healthcare often leads clinicians to perceive AI as a non-transparent entity (often referred to as the "black box" nature of AI/ML models), fostering a significant "gap of trust" that impedes the routine application of AI/ML- driven Clinical Decision Support (CDS) systems in clinical workflows. While eXplainable Artificial Intelligence (XAI) provides a pathway for mitigation, current

approaches are frequently characterized by subjectivity, highlighting a critical need for objective metrics. To facilitate impactful clinical adoption, we target the current absence of standardized, quantifiable XAI benchmarks specifically tailored for medical decision support. We aim to provide explainability at the AI-feature level by establishing novel objective metrics for AI model explainability alongside existing medical domain knowledge to make AI-based personalized medicine comprehensible and more trustworthy for physicians. To this end, we leverage Information Theory concepts such as Akaike Information Criterion (AIC), Mutual Information (MI), and Human-Centered AI. We introduce two novel objective metrics for explainability of AI models, namely Features Mutual Information (FMI) and eXplainable Information Criterion (XIC). As a metric to quantify the shared information between the features an AI model selects and those a human expert chooses which can be expressed as a measure of mutual information, FMI denotes feature level explainability by quantifying degrees of similarity of features that are used for AI model building on the one hand and in gold standard clinical practice on the other hand. Subsequently, the weighted mean of the calculated FMI and an arbitrary performance metric will be quantified to provide an objective yet task-specific measure for the explainability-performance trade-off: XIC. In conclusion, our proposed concept holds promise to bridge the gap of trust with introducing novel objective metrics to quantify AI model explainability for clinical decision support at feature level. In the future, we aim atproof of concept studies as well as systematic empirical clinician interviews to validate the usability of the proposed methodology, showcasing in post COVID diseases.

Bridging Ethical and Technical Evaluation in Machine Learning for Clinical Decision Support Systems: A Data-Centric Perspective

11:26-11:33

Presenter/Author: Sheikh Faisal Rashid

Affiliations: German Research Center for Artificial Intelligence DFKI, Berlin

The rapid integration of Machine Learning (ML) into Clinical Decision Support Systems (CDSS) has significantly advanced predictive capabilities in healthcare, including disease diagnosis, prognosis, and risk stratification. These systems increasingly rely on complex clinical datasets derived from electronic health records, medical imaging, and real-world patient data. However, recent research highlights that the performance gains of ML models often obscure critical challenges related to data quality, bias, and ethical accountability. This workshop contribution focuses on the intersection of ML development and clinical data, emphasizing the need for integrated evaluation frameworks that jointly address technical robustness and ethical validity. From a data-centric perspective, biases in clinical datasets arising from sampling disparities, missingness, or historical inequities can propagate through model development and lead to systematically unfair outcomes. Beyond traditional statistical considerations, emerging literature stresses the importance of ethical bias and ethical noise as distinct constructs that must be explicitly measured and mitigated during ML lifecycle development. Regulatory frameworks such as the European AI Act further reinforce the necessity of transparent evaluation metrics, particularly for high-risk applications like CDSS, where patient safety and

fairness are paramount. This workshop segment will present methodological approaches for identifying, quantifying, and mitigating bias and noise in clinical datasets used for ML development. It will also explore strategies such as data auditing, fairness-aware learning, and post-hoc explainability techniques to ensure compliance with ethical and regulatory standards. By integrating insights from recent clinical AI research and real-world CDSS implementations, this work aims to provide a comprehensive framework for responsible ML deployment in healthcare.

Overlooked sample quality imbalance between patient groups impairs reproducibility of clinically relevant RNA-seq studies

11:40-11:47

Presenter/Author: Maximillan Sprang

Affiliations: Dept of Dermatology, University Medical Center Mainz, Mainz

Reproducibility remains a persistent concern in biomedical research, and current publication guidelines have not solved it. While batch effects have received considerable attention, an equally important confounder is largely ignored: the imbalance in sample quality between groups being compared. In a stringently curated collection of 40 clinically relevant human RNA-seq datasets, we quantified this Quality Imbalance (QI) using machine-learning derived low-quality probabilities from our seqQscorer tool [1]. Fourteen datasets (35%) showed a high QI index (>0.30), and a comparable rate was observed for ChIP-seq. QI has a direct and measurable impact on differential gene analysis [2]. In equally sized subsets of the three largest datasets, the number of differential genes rose linearly with QI, with on average 1,222 additional genes called between perfectly balanced and perfectly imbalanced subsets. Across the 40 full datasets, the number of differentially expressed genes grew almost five times faster with sample size in highly imbalanced datasets than in balanced ones. Comparing expression with sample quality across the 13 most balanced datasets, we identified 7,708 low-quality and 5,243 high-quality marker genes recurring in at least two datasets, some in up to 77%. Low-quality markers are enriched in mitochondrial, ribosomal and stress-response pathways and are regulated by transcription factors, while high-quality markers are predominantly miRNA targets. Critically, quality markers can constitute up to 22% of the top 500 differential genes, while the proportion of known disease genes drops with rising QI, suggesting that some literature disease markers may in fact be quality artefacts. Filtering quality outliers and including quality as a confounder in the differential analysis substantially reduces these biases. Together with our earlier work on data-driven QC guidelines [3] and on quality-aware batch effect detection [4], the findings argue for routine assessment of sample-level quality alongside conventional QC and batch handling to improve reproducibility of clinically relevant sequencing studies. [1] Albrecht et al., Genome Biology 22:75 (2021). [2] Sprang et al., Genome Biology 25:222 (2024). [3] Sprang et al., Life Science Alliance 4:e202101113 (2021). [4] Sprang et al., BMC Bioinformatics 23:279 (2022).

Navigating the Ethical Landscape of Artificial Intelligence in Clinical Applications: The Role of Algorithmic Impact Assessment Tool

11:47-11:54

Presenter/Author: Srushhti Trivedi

Affiliations: University of Göttingen, Göttingen

The integration of Artificial Intelligence (AI) specially Machine Learning (ML) into medical practice holds transformative potential for enhancing diagnostic accuracy, personalizing therapeutic interventions, and optimizing healthcare delivery. However, this deployment introduces complex ethical, legal, and societal challenges demanding robust governance. This work examines the primary ethical tensions inherent in medical AI, focusing on algorithmic bias, the black box nature of deep learning models, and potential impacts on patient autonomy. It emphasizes the importance of understanding how biased training datasets can perpetuate systemic health disparities and highlights specific risks associated with insufficient transparency in AI-based clinical decisions. A crucial component of responsible AI implementation is proactive ethical evaluation. We highlight the importance of Algorithmic Impact Assessment (AIA) as a structured process for identifying, assessing, and mitigating potential harms associated with AI systems. Canada has developed and continues to implement the AIA tool [1] to promote the responsible use of AI. Germany's draft AI Act also mandates AIAs for high-risk applications, including those in healthcare, and emphasizes transparency, fairness, and accountability. While these assessments are relevant for ensuring compliance with data protection regulations (e.g., the GDPR) and for preventing discriminatory outcomes, a comparable standardized tool is currently unavailable in Germany. This requires a commitment to rigorous validation across diverse datasets, ongoing monitoring for bias, and transparent reporting of algorithmic performance. Furthermore, maintaining human oversight and ensuring patient understanding of AI-driven recommendations are crucial. Lastly, successful and equitable clinical adoption of AI depends not only on technical efficacy but on a commitment to ethical principles, proactive risk assessment through tools like AIA, and a focus on patient safety and well-being. [1]

13:30–15:00

Sponsor Workshop: SAS Data Maker + Friedrich Wingert Stiftung

Tue, 29.09.26, 13:30–15:00, Room: N4

Chair: Arpit Jain, André Sander

Authors: Arpit Jain; Sarah Gauges

Affiliations: SAS Institute, Frankfurt am Main

General outline and objectives

Drug development relies on a limited volume of relevant clinical data which is restricted in use, costly to obtain and often insufficient and underrepresented. Augmenting real-world data with synthetic data accelerates drug development through lower costs and faster execution. Synthetic data surfaces valuable trends latent in patient data while avoiding compromise of privacy. This session provides an overview of considerations for robust and governable synthetic data generation, using a new SAS offering, Data Maker as an example. It provides an overview of available methods, such as Generative Adversarial Networks (GANs), SMOTE, and other emergent methods. We also cover quality assessment and evaluation of synthetic data prior to application. We also highlight potential risks from synthetic data and suggest how to increase acceptance in drug development. The session also touches upon the stance of health authorities towards synthetic data in clinical trials, particularly in the area of drug development.

Structure

Discussing the need of Synthetic Data followed with demonstration of the SAS Data Maker solution, using real world data.

Friedrich-Wingert-Stiftung: Scholarship Holders Present Their Work

14:00-15:00

Presenter/Author: Sven Hoffmann

Affiliations: ID GmbH, Berlin

General outline and objectives

To address the shortage of qualified young talent in the field of IT within the healthcare sector, the Friedrich Wingert Foundation has developed a funding scheme to support excellent students and early-career researchers at colleges, universities and other academic institutions. Up to 10 one-year scholarships are awarded annually. In this session the awardees of the 2026 round of scholarship present their work in short talks.

Format of workshop

Short Talks - Scholarship winners present their work.

Wed 30.09.2026

09:00–10:30

AI-driven Therapeutic Modelling in Biomedical Informatics (1/2)

Wed, 30.09.26, 09:00–10:30, Room: K1

Chair: Tim Beißbarth, Ulrich Sax

Authors: Benjamin Löhnhardt [1]; Tim Beißbarth [2]; Ulrich Sax [1]; Katarzyna Bozek [3]; Adam Skubala [4]; Niko Beerenwinkel [5]; Mark Robinson [6]

Affiliations: [1] University Medical Center Göttingen, Department of Medical Informatics, Göttingen; [2] University Medical Center Göttingen, Department Medical Bioinformatics, Göttingen; [3] University of Cologne, Cologne; [4] Bayer AG, Wuppertal; [5] ETH Zurich, Basel; [6] ETH Zurich, Zurich

General outline and objectives

The workshop focuses on the application of artificial intelligence (AI) for predicting and developing new therapeutic approaches in a translational context. With the increasing availability of complex biomedical data, AI-supported methods are opening up new avenues for drug prediction, drug design and the modelling of individual therapeutic responses. The workshop will examine the topic from the perspective of academia and industry, as well as from the interdisciplinary perspective of biomedical informatics.

The workshop is organized by the GMDS-AG Biomedical Informatics and will also be a meeting of the working group. New participants for the GMDS working group on biomedical informatics are explicitly welcome.

Structure

09:00 - 09:10 Introduction (Tim Beißbarth)

09:10 - 09:45 Katarzyna Bozek: Modeling of sleep and behaviour data with machine learning

09:45 - 10:20 Niko Beerenwinkel: Modeling tumor progression from single-cell sequencing data

---

Modeling of sleep and behaviour data with machine learning

09:10-09:45

Presenter/Author: Katarzyna Bozek

Affiliations: University of Cologne, Cologne

In this talk I will present two different machine learning approaches to analysis of temporally resolved biomedical data. The first one is based on engineered features and applied to detection of

REM sleep disorder - a major prodromal marker of Parkinson's disease - from actimeter data[1]. The second one is applied to quantification of animal motion using their skeleton keypoints' trajectories and is based on fully self-supervised representation learning using Transformers[2]. Via these examples I will show the potential of machine learning in the quantification of unstructured temporal data, such as sleep or behavior recordings. Both feature engineering and self-supervised representation learning offer powerful means to extracting from the data diagnostic signals as well as detecting major motion patterns across species, conditions, and experiments. [1] Bertram et al. arXiv:2511.05221 npj Digital Medicine, accepted [2] Rose et al. Nat Methods (2026). https://doi.org/10.1038/s41592-025-02893-y

---

Modeling tumor progression from single-cell sequencing data

09:45-10:20

Presenter/Author: Niko Beerenwinkel

Affiliations: ETH Zurich, Basel

Cancer progression is a somatic evolutionary process characterized by the accumulation of genetic alterations and responsible for tumor growth, clinical progression, and drug resistance development. I will present probabilistic models for reconstructing the evolutionary history of a tumor from single-cell sequencing data, including efficient inference algorithms for mutation calling and for learning tumor phylogenies from mutation and copy number data. I will also discuss computational methods for finding common patterns of tumor evolution among patients and for inferring the underlying cellular fitness landscape of cancer evolution.

Benchmarking AI scientists

Wed, 30.09.26, 09:00–10:30, Room: K2

Chair: Georg Fuellen

Authors: Georg Fuellen [1]; Klaus Jung [2]; Sebastian Lobentanzer [3]; Christoph Bobrowski [4]; Sven De Pourcq [5]; Luis Nagai [5]

Affiliations: [1] Universitätsmedizin Rostock, Institut für Biostatistik und Informatik in Medizin und Alternsforschung, Rostock; [2] Stiftung Tierärztliche Hochschule Hannover, Hannover; [3] Helmholtz, Munich; [4] Johanniterkrankenhaus Stendal, Stendal; [5] Universitätsmedizin Rostock, Rostock

General outline and objectives

The "Benchmarking AI scientists" workshop shall give an overview on the use of AI Agents for Science, with a focus on benchmarking their performance, and a focus on the application area of biodata analysis / bioinformatics it is organized by the recently formed Projektgruppe "Künstliche Intelligenz in der Bioinformatik und Systembiologie", see https://www.gmds.de/index.php?id=2268, where also some background material on the topic can be found.

Structure

9:00 - 9:20 Sebastian Lobentanzer: Beyond Accuracy: LLM-Judge Benchmarking of AI Scientists

9:20 - 9:40 Luis Nagai, Georg Fuellen: Automated grant- and thesis-writing: how to benchmark AI-based ideation and biodata processing?

9:40 - 10:00 Christoph Bobrowski: Future research in AI desirable from the perspective of clinical medicine

10:00 - 10:30

Discussion

Beyond Accuracy: LLM-Judge Benchmarking of AI Scientists

09:00-09:20

Presenter/Author: Sebastian Lobentanzer

Affiliations: Helmholtz Zentrum München - Deutsches Forschungszentrum für Gesundheit und Umwelt (GmbH), Munich

AI agents are entering biological data analysis at remarkable speed, but our ability to measure whether they are actually useful for science has not kept pace. Standard accuracy metrics break down once tasks involve open-ended reasoning, multi-step pipelines, or domain knowledge that resists tidy multiple-choice formats. In this talk I present an extension of the BioChatter framework that addresses this evaluation gap through ground-truth-anchored LLM-as-judge benchmarking, and I situate the approach within emerging community standards for sharing benchmarks. The conceptual

core is what we call the Karenina principle, after Tolstoy's opening line, with a twist: in science, all correct and wrong answers can be correct and wrong in multiple ways. A useful benchmark therefore needs to characterise the structure of failure, not just count correct outputs. Our LLM-judge pipeline operationalises this by decomposing each ground-truth answer into rubric-style checkpoints - required claims, forbidden claims, structural requirements, and reasoning constraints - that a judge model evaluates independently. This yields finer-grained, reproducible scores than simple correctness, and makes the resulting numbers interpretable to domain experts. To make these benchmarks portable across teams, models, and time, we adopt and extend the Croissant ML metadata standard. Croissant ML provides a machine-readable description of datasets Croissant Tasks, currently being shaped within the same community effort, frames the full ML process from problem statement to expected solution. Encoding our benchmarks in Croissant Tasks turns each evaluation into a self-describing artefact: input data, task definition, scoring rubric, and reference answers travel together, are versionable, and can be re-run by any agent. This is what allows benchmarking AI scientists to scale beyond bespoke leaderboards. I will demonstrate the pipeline on biodata-analysis tasks, discuss what the Karenina-style judging reveals about current agent failure modes, and argue that ground-truth rubrics paired with Croissant-style metadata are a realistic path toward trustworthy evaluation of AI scientists in bioinformatics.

Automated software-, grant- and thesis-writing: how to benchmark AI-based ideation and biodata processing

09:20-09:40

Presenter: Georg Fuellen, Luis Nagai

Authors: Georg Fuellen; Luis Nagai; Sven De Pourcq; Anton Kulaga

Affiliations: Universitätsmedizin Rostock, Institut für Biostatistik und Informatik in Medizin und Alternsforschung, Rostock

The integration of artificial intelligence into academic and technical workflows requires rigorous benchmarking to ensure accuracy and scientific integrity. This talk explores the potential and limitations of AI-based ideation and biodata processing through preliminary results from three distinct projects. First, we examine the generation of draft German Research Foundation (DFG) Forschungsgruppenantrag proposals, focusing on the strictest-possible validation of all sections and an in-depth AI benchmarking exercise for drafting themed state-of-the-art texts. Second, we discuss the automated generation of PhD-level bioinformatics theses, emphasizing mechanisms to validate automated data analyses and accurately substantiate scientific novelty claims. Finally, we explore human-AI co-creation of software, describing how ideas from formal specification work and vericoding may enhance the reliability of AI-generated code. Ultimately, this presentation offers a pragmatic methodology for benchmarking and using AI scientist tools, ensuring that automated scientific outputs remain verifiable, precise, and rigorous.

Future research in AI desirable from the perspective of clinical medicine

09:40-10:00

Presenter/Author: Christoph Bobrowski

Affiliations: Johanniterkrankenhaus Stendal, Stendal

After a clinical example of type 1 ("heuristic") and type 2 ("analytical") reasoning, the necessity for "neuro-symbolic AI" is reviewed and the literature is reviewed. Cognitive Dual-Process Theory will be presented and their applicability in clinical medicine will be shown. Modes and methods of clinical reasoning are presented, with examples for a non-medical audience. Current AI approaches mimicing cognitive dual-process reasoning are reviewed. Arguably, deployment and acceptance of conventional clinical information systems (CIS) has been hindered by the fact that differing expectations of users and developers were never fully aligned. While this only is a matter of efficiency (and of clinician burnout!) in CIS, trust issues have arisen regarding even conventional internet based information. We believe that this gap even widens with modern AI applications in clinical medicine. Future research should focus on (i) explainabilty of results (ii) understandabilty of principles of neurosymbolic computing (iii) metrics for clinician trust (iv) integrating the well-known PDCA (plan-do-check-act) cycle which is present even in clinical patient micro-management into AI-driven clinical support systems. [1 Croskerry P. From mindless to mindful practice: cognitive bias and clinical decision making. N Engl J Med. 2013368(26):2445-2448. [2] Groopman J. How doctors think: with a new afterword. New York: Mariner Books 2008. [3] Chen MJ, Chang YN. Neuro-symbolic agentic framework for decoupled perception and clinical reasoning in medical AI. In: Proceedings of the 2026 International Conference on Advances in Artificial Intelligence and Machine Learning (AAIML) 2026 Mar Tokyo, Japan. IEEE 2026. [4] Gao Y, Li R, Croxford E, Caskey J, Patterson BW, Churpek M, et al. Leveraging medical knowledge graphs into large language models for diagnosis prediction: design and application study. JMIR AI. 20254:e58670. doi:10.2196/58670. [5] Salimparsa M, Sedig K, Lizotte DJ, Abdullah SS, Chalabianloo N, Muanda FT. Explainable AI for clinical decision support systems: literature review, key gaps, and research synthesis. Informatics. 202512(4):119. doi:10.3390/informatics12040119. [6] Sinha S, Premsri T, Kamali D, Kordjamshidi P. Toward a clearer characterization of neuro-symbolic frameworks: a brief comparative analysis. In: Proceedings of Machine Learning Research. Vol 284. Proceedings of the 19th Conference on Neurosymbolic Learning and Reasoning (NeSy 2025) 2025. p. 1-27.

Forum Digital Medicine and Artificial Intelligence (1/4)

Wed, 30.09.26, 09:00–10:30, Room: K3

Chair: Hans-Ulrich Prokosch

Authors: Hans-Ulrich Prokosch [1]; Christian Hinske [2]; Rainer Röhrig [3]; Felix Balzer [4]; Christiane Schewe [5]; Alanna Ebigbo [6]; Eimo Martens [7]; Bjoern Eskofier [8]; Daniel Röder [9]; Isabella Wiest [10]; Katharina Danhauser [11]; Hartmuth Nowak [12]; Julian Varghese [13]

Affiliations: [1] Universität Erlangen, Erlangen; [2] Lehrstuhl für Datenmanagement und Clinical Decision Support, Universität Augsburg, Augsburg; [3] Universitätsklinik RWTH Aaachen, Aachen; [4] Charité - Universitätsmedizin Berlin, Berlin; [5] Klinik und Poliklinik für Anästhesiologie, Universitätsmedizin Rostock, Rostock; [6] Klinik für Innere Medizin, Katholisches Klinikum Bochum, Bochum; [7] Klinik und Poliklinik für Innere Medizin I, University Hospital, Technical University of Munich, TUM School of Medicine and Health, Munich, Munich; [8] Lehrstuhl für KI-unterstützte Therapieentscheidungen, Ludwig-Maximilians-Universität Münch, München; [9] Anästhesiologischen Intensivstation, Universitätsklinikum Würzburg, Würzburg; [10] Else Kröner Fresenius Zentrum für Digitale Gesundheit TU Dresden, Dresden; [11] LMU Munich, München; [12] Universitätsklinikum Knappschaftskrankenhaus Bochum, Bochum; [13] Institute of Medical Data Science, Otto von Guericke University Magdeburg, Magdeburg

General outline and objectives

The workshop is part of the Forum Digital Medicine and Artificial Intelligence, which aims to bring together Clinformaticians (Clinical Informaticians) from the clinical societies of the AWMF with the methodological disciplines of the GMDS in order to discuss interdisciplinary methodological, regulatory, and political challenges in digital medicine as well as for the development and application of Artifical Intelligence Tools. Based on stimulating expert presentations the workshop participants will discuss the current challenges and hurdles and interactively develop solutions for future developments and activities. One part of the workshop will be a Pro&Contra Presentation and Discussion on the opportunities, but also the challenges and hurdles for agentic AI systems.

Structure

09:00 - 10:30 Workshop-Part 1

09:00 - 09:30 Felix Balzer: Hospital Information System 4.0 - What will the next generation of HIS look like?

09:30 - 10:00 Christiane Schewe: CMIO Job Description: What exactly is it, and what does it take to become one?

10:00 - 10:30 Alanna Ebigbo: Clinical Use of AI in Endoscopy - Is there real clinical value?

Hospital Information System 4.0 - What will the next generation of HIS look like?

09:00-09:30

Presenter/Author: Felix Balzer

Affiliations: Charité - Universitätsmedizin Berlin, Berlin

Während klassische KIS primär Dokumentation und Abrechnung abbilden, entsteht mit "KIS 4.0" eine neue Generation datengetriebener Plattformen. Im Zentrum stehen Architekturen, die Versorgung und Forschung nicht mehr trennen, sondern klinische Routinedaten in Echtzeit für Entscheidungsunterstützung, Prozesssteuerung und wissenschaftliche Nutzung verfügbar machen. Der Vortrag ordnet den aktuellen Entwicklungsstand ein und diskutiert anhand konkreter Beispiele zentrale Designentscheidungen sowie Herausforderungen auf dem Weg zur nächsten KIS-Generation.

CMIO Job Description: What exactly is it, and what does it take to become one?

09:30-10:00

Presenter/Author: Christiane Schewe

Affiliations: Klinik und Poliklinik für Anästhesiologie, Universitätsmedizin Rostock, Rostock

Test von Ulli Prokosch - muss noch korrigiert werden Test2 um zu sehen, ob ich den Abstract noch mal ändern kann

---

Clinical Use of AI in Endoscopy - Is there real clinical value?

10:00-10:30

Presenter/Author: Alanna Ebigbo

Affiliations: Klinik für Innere Medizin, Katholisches Klinikum Bochum, Bochum

(german version below) In gastrointestinal endoscopy, AI research focuses on the detection and characterization of early-stage cancers and precancerous lesions. Commercially available products with CE marking are currently limited to the detection and characterization of adenomas during screening colonoscopy, as well as the detection of dysplasia and early-stage cancers in Barrett's esophagus. This presentation will focus on the diagnostic performance of these systems in clinical decision support, as well as future developments in the use of AI in gastrointestinal endoscopy. In der gastrointestinalen Endoskopie liegt der Schwerpunkt der KI-Forschung auf der Detektion und Charakterisierung von Frühkarzinomen und Präkanzerosen. Kommerziell erhältliche Produkte mit CE-Zulassung beschränken sich derzeit auf die Detektion und Charakterisierung von Adenomen im Rahmen der Vorsorgekoloskopie sowie auf die Detektion von Dysplasien und Frühkarzinomen beim Barrett-Ösophagus. Die diagnostische Leistungsfähigkeit dieser Systeme in der klinischen Entscheidungsunterstützung sowie zukünftige Entwicklungen des KI-Einsatzes in der gastrointestinalen Endoskopie werden im Mittelpunkt dieses Vortrags stehen.

Regulatory Topics for Statisticians: Implementation of Risk-Based Quality Management from Biostatistical Perspective

Wed, 30.09.26, 09:00–10:30, Room: K4

Chair: Anika Großhennig

Authors: Hans Bauer [1]; Hannes Buchner [1]; Antonia Zapf [2]; Anja Sander [3]; Anika Großhennig [4]; Maren Vens [5]

Affiliations: [1] Staburo GmbH, München; [2] Department of Medical Biometry, University Medical Center Hamburg-Eppendorf, Hamburg; [3] Universitätsklinikum Heidelberg, Heidelberg; [4] Medizinische Hochschule Hannover, Hannover; [5] Universität zu Lübeck, Institut für Medizinische Biometrie und Statistik, Universitätsklinkum Schleswig-Holstein, Campus Lübeck, Lübeck

General outline and objectives

To ensure data quality, patient safety, and operational efficiency, the implementation of Risk-Based Quality Management (RBQM) has become essential in clinical trials. As emphasized in the ICH E6(R3) guideline, RBQM is a critical framework for safeguarding the integrity of clinical trial data. By systematically identifying, assessing, and prioritizing risks based on their potential impact, RBQM enables targeting monitoring strategies that enhance regulatory compliance, accelerates study execution, and strengthen stakeholders' confidence in trial outcomes.

In this workshop, we will explore practical approaches to implementing RBQM in clinical research and highlight the role that study biostatisticians can or should play in this context. In the workshop, classical strategies, such as central monitoring of missing data, detection of adverse event under- or over-reporting, and trends in protocol deviations, will be discussed. In addition, the workshop will place special emphasis on advanced data analytics including comprehensive database-wide analyses for early issue and misbehaviour detection. Participants will gain insights into leveraging statistical and data-driven tools to proactively identify risks, improve quality of relevant data, and optimize monitoring efforts across multi-site trials.

Structure

9:00 - 9:10 Hannes Buchner: Introduction into Risk-Based Quality Management (RBQM)

9:10 - 9:40 Maren Vens: Best RBQM practice for investigator-initiated trials (IITs)

9:40 - 10:10 Hans Bauer: Whole-database analysis for central monitoring and RBQM

10:10 - 10:30

Discussion with discussant Anja Sander and speakers

Introduction into Risk-Based Quality Management (RBQM)

09:00-09:10

Presenter/Author: Hannes Buchner

Affiliations: Staburo, Munich

This presentation will provide an overview of the principles and significance of Risk-Based Quality Management (RBQM). The increasing complexity of clinical studies, escalation of costs and the growing international need for more effective quality management systems led to the development of RBQM. The guideline ICH E6(R3) - effective since July 2025 - focusses on RBQM and "Quality by Design" (QbD) for clinical studies. This concept has been developed internationally and accepted by FDA. Key elements of RBQM are: Proactive Quality Management = Quality by Design (not just retrospective controls), focus on critical data (CTQ = critical-to-quality factors), and risk-based surveillance including centralized monitoring. The impact on CTQs is to prioritize areas that directly impact patient safety, data reliability or trial integrity. KRI's (key risk indicators) and definition of QTL's (Quality tolerance limits) in conjunction with data surveillance are essential parameters for any quality management. A good risk management strategy (including risk identification, risk assessment, risk control, risk communication and risk monitoring) is fundamental for a successful clinical trial in all stages of a clinical study, from planning & design to execution and finally the documentation and resolution stage. RBQM is cross-functional and facilitates communication among all involved parties and clear escalation pathways to maintain effective oversight and management of the identified risks. The good use of RBQM will ensure data integrity and patient safety across all stages of a clinical study.

Best RBQM practice for investigator-initiated trials (IITs)

09:10-09:40

Presenter/Author: Maren Vens

Affiliations: Universität zu Lübeck, Institut für Medizinische Biometrie und Statistik, Universitätsklinkum Schleswig-Holstein, Campus Lübeck, Lübeck

In investigator-initiated trials (IITs), resources are often limited, the trial infrastructure varies, and coordination between different institutions is complex. Nevertheless, ensuring data quality, patient safety, and regulatory compliance is just as critical as in industry-sponsored trials. In this presentation, we will examine existing challenges and highlight concrete, practical approaches for successfully implementing RBQM in IITs. Using real-world examples from ongoing and completed IITs, we explain how early risk detection and key monitoring strategies can be implemented in practice from a biostatistical point of view. We demonstrate how risk profiles for IITs were developed, critical risks were prioritized, and monitoring activities were strategically managed using biostatistical input. These examples illustrate how monitoring can be supported, potential problems identified early on, and communication between study partners improved. The presentation offers concrete, proven recommendations for effectively and sustainably implementing RBQM, even under the unique conditions of investigator-initiated trials.

Whole-database analysis for central monitoring and RBQM

09:40-10:10

Presenter/Author: Hans Bauer

Affiliations: Staburo GmbH, München

Central monitoring enables Risk-Based Quality Management by quantifying risks related to sites (or other entities) involved in a clinical study, allowing for targeted follow-up activities. However, the methods used to quantify risks in this context can vary widely in the amount of information utilised, ranging from comparison of data points for individual patients to aggregating all collected data into a single number to compare between sites. Assuming (importantly) a valid underlying methodology, the distinctive potential of central monitoring via data analytic and statistical techniques is in detecting signals discernible only across patients and database tables, or even across the entire database. The complexity of data analysis, however, increases considerably with the number of database tables to be assessed in conjunction. In this talk, a "whole-database" approach to data analysis will be outlined which essentially considers the entire clinical database as a single table, enabling for example the computation and subsequent aggregation of a large number of statistical comparisons in a "content-agnostic" manner to detect data anomalies that would be hard or impossible to discern otherwise, as well as other cross-table operations such as verification of data relations across tables (e.g. tumour lesion and response data), derivation of global data quality metrics, or complex data normalisation (e.g. generation of a unified lab dataset). The potential of using R for implementing such an approach and the associated implementation challenges will be discussed and illustrated based on non-confidential (dummy or publicly available) data.

Ordinal Outcomes: Friend or Foe?

Wed, 30.09.26, 09:00–10:30, Room: Runder Saal

Chair: Katherine Lee

Authors: Katherine Lee [1]; Leon Di Stefano [1]; Chris Selman [2]; Johan Verbeeck [3]; Ian White [4]

Affiliations: [1] Murdoch Children's Research Institute, Melbourne; [2] Murdoch Children's Research Institute, Parkville; [3] Hasselt University, Hasselt; [4] MRC CTU at UCL, London

General outline and objectives

Ordinal outcomes are becoming increasingly popular in the medical and public health literature, particularly in platform trials where it is desirable to have a common primary outcome that is relevant across heterogeneous groups of participants and captures a wide range of clinical outcomes. This workshop will involve a series of presentations followed by a facilitated discussion on the development, evaluation and analysis of ordinal outcomes, which we hope will spark a discussion of whether ordinal outcomes are our friend or foe.

Structure

9:00 - 9:15 Leon Di Stefano: Constructing and communicating ordinal outcomes

9:15 - 9:30 Chris Selman: Evaluating novel ordinal outcomes for use in randomised controlled trials

9:30 - 9:45 Johan Verbeeck: Hierarchical composite endpoints: GPC, DOOR, and MOST as ordinal frameworks to capture patient-centered outcomes

9:45 - 10:00 Ian White: The meaning of model-based estimands for ordinal outcomes

10:00 - 10:30

Panel Discussion

Constructing and communicating ordinal outcomes

09:00-09:15

Presenter/Author: Leon Di Stefano

Affiliations: Murdoch Children's Research Institute, Melbourne

Ordinal outcomes are an increasingly popular choice for clinical trials. This talk will discuss several issues pertinent to constructing these outcomes and communicating them. The first is how ordinal outcomes can be built up from existing components. We identify three broad mechanisms from the literature: (1) overriding, (2) complete tie-breaking, and (3) what we term "parallel composition", which has a number of forms. We illustrate each of these with examples and discuss their strengths and weaknesses. The second issue is how many levels an ordinal outcome should have. We provide additional analysis of the asymptotic results of Whitehead (1993) to show that statistical efficiency

is strongly constrained by the largest category probability, and discuss the implications of this for the tradeoffs around more versus fewer categories. The third issue is how (comparisons of) ordinal outcome distributions can be communicated and graphically displayed. We discuss alternatives including stacked bar charts, eCDF and survival plots, and PP plots and assess these from the perspective of the grammar of graphics, the Cleveland-McGill hierarchy, and clinical decision-making principles. Finally, we argue for the utility of clearly delineating the task of constructing a composite ordinal outcome from the related but distinct tasks of evaluating the outcome using pre-existing data, choosing an effect measure, choosing an estimator, and handling censoring. Whitehead J. Sample size calculations for ordered categorical data. Statistics in Medicine. 199312(24):2257-2271. doi:10.1002/sim.4780122404

Evaluating novel ordinal outcomes for use in randomised controlled trials

09:15-09:30

Presenter: Chris Selman

Authors: Chris Selman [1]; Katherine Lee [2, 3]; Ian White [4]; Sean Ong [5]; Melissa Middleton [2]; Leon Di Stefano [2]; Anneke Grobler [2]; Hannah Walker [2]; Rachel Conyers [2]; Robert Mahar [3]

Affiliations: [1] Murdoch Children's Research Institute, Parkville; [2] Murdoch Children's Research Institute, Melbourne; [3] University of Melbourne, Melbourne; [4] UCL, London; [5] Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto

Introduction

Selecting an appropriate primary outcome is critical in randomised controlled trials (RCTs) since it guides trial design and analysis and ensures the clinical relevance of conclusions. In recent years, ordinal outcomes have become common because they can capture multiple patient states and improve statistical efficiency compared to binary outcomes. This has led to many novel ordinal outcomes being developed for use in RCTs. It is important that novel ordinal outcomes be rigorously evaluated to ensure clinical relevance, interpretability, and reliability prior to use, although there is substantial variation in how this is currently being done. In this talk, we aim to provide practical guidance on evaluating novel ordinal outcomes.

Methods

In this talk, we describe a prospective framework for evaluating the utility of novel ordinal outcomes for use in an RCT. This framework is illustrated using four ordinal outcomes that were developed and evaluated as candidates for use in a trial in paediatric patients receiving haematopoietic stem cell transplant.

Results

We propose and discuss an evaluation framework with six elements: (1) obtain evaluation data, (2) assess the distribution of participants across the outcome categories, (3) assess clinical validity of the outcome via expert review of patient vignettes, (4) investigate effects with known interventions or prognostic factors, (5) estimate associations with established outcomes, and (6) assess assumptions required for analysis. The worked example demonstrates the feasibility and utility of this framework.

Conclusion

Evaluating novel ordinal outcomes requires a comprehensive approach that assesses both statistical performance and clinical relevance. We hope that providing a structured framework in this talk will help researchers evaluate novel ordinal outcomes that are meaningful for trial decision-making.

Hierarchical composite endpoints: GPC, DOOR, and MOST as ordinal frameworks to capture patient-centered outcomes

09:30-09:45

Presenter/Author: Johan Verbeeck

Affiliations: Hasselt University, Hasselt

Ordinal outcomes offer a compelling framework for capturing the breadth of a patient's clinical experience in a single, interpretable measure. Hierarchical composite endpoints join multiple outcomes in a pre-specified clinical priority ordering, naturally resulting, in specific cases, in ordinal outcomes. This talk presents a comparative overview of three hierarchical multi-component frameworks: generalized pairwise comparisons (GPC), the desirability of outcome ranking (DOOR), and the Markov ordinal state transition model (MOST). All three construct, implicitly or explicitly, an ordinal structure over patient outcomes. DOOR and MOST are ordinal outcomes by construction: DOOR maps multiple components into an ordinal scale analysed via rank-based inference, while MOST tracks longitudinal transitions across ordinal health states. GPC, encompassing statistics such as the win ratio and net treatment benefit, operates through pairwise comparisons across a pre-specified outcome hierarchy, and can accommodate endpoints beyond settings where components are collapsed into a single ordinal scale. Through a structured exploration of their statistical and philosophical underpinnings, we highlight where GPC, DOOR, and MOST converge and where they diverge, in terms of estimands, assumptions, handling of ties and censoring, and suitability for different trial contexts, and explore their connections to classical ordinal methodology. This overview will clarify when hierarchical composite endpoints serve as genuine allies in trial design, and where methodological or practical challenges warrant caution.

The meaning of model-based estimands for ordinal outcomes

09:45-10:00

Presenter/Author: Ian White

Affiliations: UCL, London

The proportional odds model is commonly used for the primary analysis of ordered categorical outcomes in two-arm trials. Fitting this model estimates an odds ratio that is assumed common across all possible ways of dichotomising the outcome. But what is the meaning of this estimated odds ratio if the assumption is not, or may not be, true? Harrell has shown that it is closely linked to the concordance probability, the probability that a randomly chosen member of the new intervention group has better outcome than a randomly chosen member of the control group. This is turn is closely related to the win difference (a form of generalised pairwise comparison). These summaries, while being free of assumptions, reflect only the direction and not the magnitude of any difference. I show by simulation that it is possible to have a treatment that is beneficial according to the concordance probability, but is harmful in terms of the clinical utility of the various states. I discuss how plausible this conflict is and what can be done to avoid it. Key ideas are to use other outcome summaries (e.g. clinical utility or dichotomisation at a clinically important level) as sensitivity analyses.

11:00–12:30

AI-driven Therapeutic Modelling in Biomedical Informatics (2/2)

Wed, 30.09.26, 11:00–12:30, Room: K1

Chair: Tim Beißbarth, Ulrich Sax

Authors: Benjamin Löhnhardt [1]; Tim Beißbarth [2]; Ulrich Sax [1]; Katarzyna Bozek [3]; Adam Skubala [4]; Niko Beerenwinkel [5]; Mark Robinson [6]

Affiliations: [1] University Medical Center Göttingen, Department of Medical Informatics, Göttingen; [2] University Medical Center Göttingen, Department Medical Bioinformatics, Göttingen; [3] University of Cologne, Cologne; [4] Bayer AG, Wuppertal; [5] ETH Zurich, Basel; [6] ETH Zurich, Zurich

General outline and objectives

The workshop focuses on the application of artificial intelligence (AI) for predicting and developing new therapeutic approaches in a translational context. With the increasing availability of complex biomedical data, AI-supported methods are opening up new avenues for drug prediction, drug design and the modelling of individual therapeutic responses. The workshop will examine the topic from the perspective of academia and industry, as well as from the interdisciplinary perspective of biomedical informatics.

The workshop is organized by the GMDS-AG Biomedical Informatics and will also be a meeting of the working group. New participants for the GMDS working group on biomedical informatics are explicitly welcome.

Structure

11:00 - 11:35 Mark Robinson: Next generation evaluation of (AI) tools in computational biology

11:35 - 12:10 Adam Skubala: Pioneering Prognostic Score Adjustment in clinical trials: A Novel Application of Machine Learning in the clinical trial setting

12:10 - 12:30 Summary and concluding discussion (Ulrich Sax)

Workshop description: https://www.gmds.de/aktivitaeten/medizinische-informatik/arbeitsgruppen-seiten/biomedizinische-informatik-gemeinsame-arbeitsgruppe-des-fachbereiches-med-informatik-und-biometrie/workshop-2026/

Next generation evaluation of (AI) tools in computational biology

11:00-11:35

Presenter/Author: Mark Robinson

Affiliations: ETH Zurich, Zurich

As artificial intelligence becomes increasingly embedded in computational biology, the challenge is no longer only to develop new methods, but to evaluate them rigorously, fairly, and reproducibly. In this talk, I will discuss the challenges of benchmarking computational and AI-driven tools in biology, including dataset bias, shifting ground truth, biological heterogeneity, metric selection, and reproducibility across rapidly evolving methods and software environments. I will also present Omnibenchmark, a framework for reproducible, open, collaborative, and extensible benchmarking designed to support transparent method evaluation and community-driven standards in computational biology.

Pioneering Prognostic Score Adjustment in clinical trials: A Novel Application of Machine Learning in the clinical trial setting

11:35-12:10

Presenter: Adam Skubala

Authors: Adam Skubala [1]; Eliana Garcia Cossio [2]; Antigoni Elefsinioti [2]

Affiliations: [1] Bayer AG, Wuppertal; [2] Bayer AG, Berlin

Advanced statistical methodology and machine learning can strengthen clinical trial inference by improving the precision of treatment-effect estimates and the efficiency of clinical development. In this talk, Bayer shares practical experience implementing machine-learning-based prognostic score adjustment as part of pre-specified supportive analyses in Phase II/III trials. This was the first regulatory-grade implementation of this approach at Bayer, establishing a reusable, governed blueprint for subsequent programs. Prognostic score adjustment uses a baseline 'risk score' (the predicted outcome under standard of care) to adjust the primary endpoint analysis, reducing residual variability and increasing estimation efficiency. The approach trains a prognostic model on baseline covariates using historical control data, initially from completed clinical trials and, looking ahead, potentially from real-world data sources. The resulting prognostic score is then applied as an adjustment variable in a pre-specified supportive (sensitivity) analysis that runs alongside, and does not alter, the primary efficacy analysis. Because the prognostic score is pre-specified and derived solely from external control data, the adjustment is designed to preserve the validity of the primary inference whether or not it improves precision. Machine learning is well-suited to building such scores because it flexibly captures nonlinear relationships and interactions among prognostic factors while maintaining strong predictive performance. Although our first implementation used an established, internally accepted method, the framework is intentionally model-agnostic: any algorithm, including newer foundation-model approaches for tabular data, may be used if it meets predefined, fit-for-purpose criteria for performance, stability, and validation. To support interpretability in cross-functional and regulatory-facing review, we incorporate explainable AI (xAI) outputs that clarify which baseline factors drive predictions. A key enabler was an end-to-end, GxP-compliant (auditable and reproducible) analysis pipeline, built through close collaboration across Bayer's data science, biostatistics, statistical programming, and statistical methodology teams. It supports reproducible execution across programming languages and platforms and follows best practices for pre-specification, separation of

training and analysis, documentation, and review. Overall, this work shows how prognostic modelling and xAI can be integrated into regulated clinical trial workflows to deliver more precise supportive evidence, while building the operational experience needed for broader adoption and constructive engagement with regulators.

Forum Digital Medicine and Artificial Intelligence (2/4)

Wed, 30.09.26, 11:00–12:30, Room: K3

Chair: Hans-Ulrich Prokosch

Authors: Hans-Ulrich Prokosch [1]; Christian Hinske [2]; Rainer Röhrig [3]; Felix Balzer [4]; Christiane Schewe [5]; Alanna Ebigbo [6]; Eimo Martens [7]; Bjoern Eskofier [8]; Daniel Röder [9]; Isabella Wiest [10]; Katharina Danhauser [11]; Hartmuth Nowak [12]; Julian Varghese [13]

Affiliations: [1] Universität Erlangen, Erlangen; [2] Lehrstuhl für Datenmanagement und Clinical Decision Support, Universität Augsburg, Augsburg; [3] Universitätsklinik RWTH Aaachen, Aachen; [4] Charité - Universitätsmedizin Berlin, Berlin; [5] Klinik und Poliklinik für Anästhesiologie, Universitätsmedizin Rostock, Rostock; [6] Klinik für Innere Medizin, Katholisches Klinikum Bochum, Bochum; [7] Klinik und Poliklinik für Innere Medizin I, University Hospital, Technical University of Munich, TUM School of Medicine and Health, Munich, Munich; [8] Lehrstuhl für KI-unterstützte Therapieentscheidungen, Ludwig-Maximilians-Universität Münch, München; [9] Anästhesiologischen Intensivstation, Universitätsklinikum Würzburg, Würzburg; [10] Else Kröner Fresenius Zentrum für Digitale Gesundheit TU Dresden, Dresden; [11] LMU Munich, München; [12] Universitätsklinikum Knappschaftskrankenhaus Bochum, Bochum; [13] Institute of Medical Data Science, Otto von Guericke University Magdeburg, Magdeburg

General outline and objectives

The workshop is part of the Forum Digital Medicine and Artificial Intelligence, which aims to bring together Clinformaticians (Clinical Informaticians) from the clinical societies of the AWMF with the methodological disciplines of the GMDS in order to discuss interdisciplinary methodological, regulatory, and political challenges in digital medicine as well as for the development and application of Artifical Intelligence Tools. Based on stimulating expert presentations the workshop participants will discuss the current challenges and hurdles and interactively develop solutions for future developments and activities. One part of the workshop will be a Pro&Contra Presentation and Discussion on the opportunities, but also the challenges and hurdles for agentic AI systems.

Structure

11:00 - 12:30 Workshop-Part 2

11:00 - 11:30 Björn Eskoffier: AI-powered digital medicine-from basic research to prescription-based products

11:30 - 12:00 Eimo Martens: Telemedicine in cardiology

AI-powered digital medicine-from basic research to prescription-based products

11:00-11:30

Presenter/Author: Bjoern Eskofier

Affiliations: Lehrstuhl für KI-unterstützte Therapieentscheidungen, Ludwig-Maximilians-Universität Münch, München

Künstliche Intelligenz (KI) entwickelt sich zunehmend zu einer neuen Basistechnologie der Medizin. Fortschritte bei Foundation Models, multimodaler Datenanalyse und agentischen KI-Systemen eröffnen neue Möglichkeiten für Diagnostik, Therapie und Patientenbegleitung entlang des gesamten Versorgungspfades. Gleichzeitig wächst die Verfügbarkeit digitaler Gesundheitsdaten - von klinischen Routinedaten über Bildgebung bis hin zu patientengenerierten Gesundheitsdaten aus Wearables, Sensorik und digitalen Gesundheitsanwendungen. Trotz dieser Dynamik gelingt die Translation KI-basierter Innovationen in die Regelversorgung bislang nur begrenzt. Ursachen sind unter anderem fragmentierte Gesundheitsdateninfrastrukturen, mangelnde Interoperabilität, regulatorische Hürden sowie die Herausforderung, aus technologischen Demonstratoren tatsächlich klinisch nutzbare und erstattungsfähige Produkte zu entwickeln. Der Vortrag beleuchtet den Weg der KI-unterstützten digitalen Medizin "von der Grundlagenforschung bis zum Produkt auf Rezept". Vorgestellt werden aktuelle Entwicklungen aus Forschung und Translation - von Personal Health Dataspaces und föderierten Datenräumen über KI-gestützte klinische Entscheidungsunterstützung bis hin zu digitalen Medizinprodukten und DiGA-Anwendungen. Anhand konkreter Beispiele aus Sensorik, Parkinson-Forschung, Digital Therapeutics und klinischer KI wird gezeigt, wie aus akademischer Forschung Versorgungslösungen mit realem Patientennutzen entstehen können. Darüber hinaus werden zentrale Herausforderungen diskutiert, darunter Evidenzgenerierung, Vertrauen und Explainability, regulatorische Anforderungen (MDR, EU AI Act, EHDS) sowie neue Rollenmodelle für Kliniken, Startups und Forschungseinrichtungen in einem zukünftigen KI-getriebenen Gesundheitssystem. Ziel ist es, Perspektiven für eine digitale Medizin aufzuzeigen, die medizinische Entscheidungen objektiver, präziser, personalisierter und skalierbarer macht.

Telemedicine in Cardiology

11:30-12:00

Presenter/Author: Eimo Martens

Affiliations: Klinik und Poliklinik für Innere Medizin I, University Hospital, Technical University of Munich, TUM School of Medicine and Health, Munich, Munich

(german version below) Cardiology faces rising numbers of chronically ill patients and worsening regional care gaps, while evidence for telemedicine has matured and, since 2022/2023, entered reimbursed routine care in Germany. This talk reviews the modalities of cardiological telemedicine and the clinical, scientific and economic case for embedding them into a networked, intersectoral infrastructure. Telemonitoring of patients with heart failure and cardiac implantable electronic devices (ICD, CRT, ICM) reduces mortality and hospitalisation: randomised (TIM-HF2) and meta-analytic data show roughly a 16% relative reduction in all-cause mortality and a 15-19% reduction in heart-failure hospitalisations. Yet routine monitoring of a large cohort generates a substantial alert load (in our >4,000-patient centre, on average ~8 alerts per patient per month), making automation indispensable. AI and signal-processing algorithms already triage remote ECGs and device episodes

with >95% accuracy and improve prediction of impending decompensation, allowing scarce specialist capacity to focus on actionable events. Drawing on work at the TUM Device Therapy and Telemedicine Centre, I illustrate how a DGK-certified, KIS- and ePA-integrated platform enables shared, resource-sparing care across hospitals, cardiology and primary-care practices, and the patient's home - extending high-quality care to remote regions, lowering costs, and reducing the CO2 footprint. Care will remain hybrid, but interoperable, AI-supported telemedicine can overcome several of today's central challenges. Die Kardiologie sieht sich mit einer steigenden Zahl chronisch kranker Patienten und zunehmenden regionalen Versorgungslücken konfrontiert, während die Evidenz für die Telemedizin ausgereift ist und diese seit 2022/2023 in Deutschland als erstattungsfähige Routineversorgung etabliert wurde. Dieser Vortrag gibt einen Überblick über die Modalitäten der kardiologischen Telemedizin sowie über die klinischen, wissenschaftlichen und wirtschaftlichen Argumente für deren Einbettung in eine vernetzte, sektorübergreifende Infrastruktur. Das Telemonitoring von Patienten mit Herzinsuffizienz und implantierbaren Herzgeräten (ICD, CRT, ICM) senkt die Mortalität und die Zahl der Krankenhausaufenthalte: Randomisierte (TIM-HF2) und metaanalytische Daten zeigen eine relative Senkung der Gesamtmortalität um etwa 16 Prozent sowie eine Verringerung der Krankenhausaufenthalte aufgrund von Herzinsuffizienz um 15-19 %. Die routinemäßige Überwachung einer großen Kohorte erzeugt jedoch eine erhebliche Alarmlast (in unserem Zentrum mit >4.000 Patienten durchschnittlich ~8 Alarme pro Patient und Monat), was eine Automatisierung unverzichtbar macht. KI- und Signalverarbeitungsalgorithmen führen bereits eine Triage von Fern-EKGs und Geräteereignissen mit einer Genauigkeit von >95 % durch und verbessern die Vorhersage einer drohenden Dekompensation, sodass sich die knappen Fachkräftekapazitäten auf handlungsrelevante Ereignisse konzentrieren können. Anhand von Arbeiten des Zentrums für Gerätetherapie und Telemedizin der TUM veranschauliche ich, wie eine DGK-zertifizierte, in KIS- und ePA-Systeme integrierte Plattform eine gemeinsame, ressourcenschonende Versorgung über Krankenhäuser, kardiologische und hausärztliche Praxen sowie das Zuhause des Patienten hinweg ermöglicht - wodurch eine hochwertige Versorgung auch in abgelegene Regionen ausgeweitet, Kosten gesenkt und der CO_2-Fußabdruck verringert werden. Die Versorgung wird weiterhin hybrid bleiben, doch eine interoperable, KI-gestützte Telemedizin kann viele der heutigen zentralen Herausforderungen bewältigen.

DGAI WAK Digital Medicine and data-driven telemedicine

12:00-12:30

Presenter/Author: Daniel Röder

Affiliations: Anästhesiologischen Intensivstation, Universitätsklinikum Würzburg, Würzburg

(german version below) The Digital Medicine Working Group (WAK DigiMed) of the German Society of Anesthesiology and Intensive Care Medicine (DGAI) is a platform for clinically active physicians jointly advancing the digital transformation in anesthesiology, intensive care, emergency, pain, and palliative medicine. This presentation introduces the vision and roadmap of the WAK. Semantic standardization, PDMS interoperability, and evidence generation are our core clinical challenges.

How can we, as Clinformaticians, move forward together with medical informaticians, biostatisticians, and data scientists? Der Wissenschaftliche Arbeitskreis Digitale Medizin (WAK DigiMed) der Deutschen Gesellschaft für Anästhesiologie und Intensivmedizin (DGAI) ist eine Plattform klinisch tätiger Ärztinnen und Ärzte, die gemeinsam die digitale Transformation in Anästhesiologie, Intensivmedizin, Notfall-, Schmerz- und Palliativmedizin gestalten. Der Vortrag stellt die Vision und die Roadmap des WAK vor. Semantische Standardisierung, PDMS-Interoperabilität und Evidenzgenerierung sind unsere zentralen klinischen Herausforderungen. Wie können wir Clinformaticians gemeinsam mit Medizininformatikerinnen, Biometriker und Data Scientists die nächsten Schritte gemeinsam gehen?

Population-Adjusted Indirect Comparisons in the Framework of Health Technology Assessment

Wed, 30.09.26, 11:00–12:30, Room: K4

Chair: Anika Großhennig, Ralf Bender, Friedhelm Leverkus

Authors: Ralf Bender [1]; Friedhelm Leverkus [2]; Anika Großhennig [3]; David Phillippo [4]; Sarah Böhme [5]; Robert Bauer [6]; Tobias Bluhmki [7]; David Mc Connell [8]

Affiliations: [1] Institute for Quality and Efficiency in Health Care, Köln; [2] Evidence-generating Data Evaluation, Berlin; [3] Medizinische Hochschule Hannover, Hannover; [4] University of Bristol, Bristol; [5] Pfizer Pharma GmbH, Berlin; [6] Novo Nordisk, Søborg; [7] Bristol Myers Squibb, München; [8] National Centre for Pharmacoeconomics, Dublin

General outline and objectives

Standard methods for indirect comparisons and network meta-analysis are based upon aggregated data and require the validity of the main assumptions of sufficient similarity, homogeneity, and consistency. The assumption of similarity is also called "constancy of relative effects", because it requires that all effect modifiers are balanced between the two trial populations. Frequently, typical questions in the HTA framework lead to indirect comparisons where the assumption of sufficient similarity is questionable or obviously invalid. For the case that individual patient data are available in one subset of trials, methods have been developed to relax the assumption of similarity by performing population-adjustments for effect modifiers, which are not balanced in the included trials. The validity of the corresponding results depends on the new assumption that all relevant effect modifiers are known and correctly included in the statistical model. This assumption is called "conditional constancy of relative effects". Methods for population-adjusted indirect comparisons have been generalized to allow comparisons in disconnected networks and the inclusion of single-arm trials. These methods, however, make the much stronger assumption of "conditional constancy of absolute effects", which means that the absolute treatment effect is constant at any level of the included variables. Thus, beside all effect modifiers, in addition all relevant prognostic variables have to be known and correctly included in the model. Not only is this strong assumption unlikely to be valid in practice, but it is furthermore extremely challenging to provide sufficient evidence that this assumption is fulfilled in a given indirect comparison.

In this workshop, methods for population-adjusted indirect comparisons (PAICs) are introduced and the advantages and disadvantages of the most important approaches are discussed. A special focus is given on multilevel network meta-regression, because this method has several advantages compared to matching-adjusted indirect comparisons and simulated treatment comparisons. The importance, suitability and limitations of PAICs in the HTA framework are discussed, along with recent methodological developments and first experiences of the new joint clinical assessments at EU level from industry, academic and HTA perspectives.

Structure

11:00 - 11:10 Ralf Bender: Indirect comparisons: Introduction and terminology

11:10 - 11:30 David Phillippo: Multilevel network meta-regression for HTA: Target estimands, assumptions, and recent developments

11:30 - 11:50 Sarah Böhme, Rober Bauer, Tobias Bluhmki: Population-adjusted indirect comparisons in health technology assessment: Practical insights from an industry perspective

11:50 - 12:10 David Mc Connell: Role of population-adjusted indirect comparisons in EU joint clinical assessments

12:10 - 12:30

Discussion

Indirect comparisons: Introduction and terminology

11:00-11:10

Presenter/Author: Ralf Bender

Affiliations: Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen, Köln

It is widely acknowledged that meta-analyses of several randomised controlled trials with low risk of bias represent the gold standard for clear-cut conclusions in the framework of health technology assessment (HTA). Frequently, typical HTA questions require indirect comparisons, which have, in general, a lower certainty of results compared to adequate direct comparisons. A variety of approaches for indirect comparisons and network meta-analyses are available for different data situations and these methods are associated with quite different uncertainties. An overview of the various approaches for indirect comparisons and network meta-analyses is presented to provide a fundamental groundwork relevant to the subsequent talks in the workshop. The main assumptions of the different approaches and the terminology for indirect comparisons are summarized. It is important to distinguish between different data situations, such as connected versus disconnected networks, and to be aware of the main assumptions behind the available methods. The network structure and the validity of the assumptions of the applied method have important consequences for the certainty of results. The presented short overview of the various approaches for indirect comparisons and network meta-analyses and their assumptions should help the workshop participants to follow the subsequent talks of the workshop.

Multilevel network meta-regression for HTA: Target estimands, assumptions, and recent developments

11:10-11:30

Presenter/Author: David Phillippo

Affiliations: University of Bristol, Bristol

Health technology assessment (HTA) requires reliable estimates of relative treatment effects, often obtained using network meta-analysis (NMA) and indirect comparison methods. These methods combine aggregate data (AgD) from multiple trials on treatments of interest, assuming that any effect modifiers are balanced across study populations and that the evidence forms a connected network of randomised trials. Population adjustment methods are increasingly used when these requirements are not met, such as matching-adjusted indirect comparison (MAIC), simulated treatment comparison (STC), and multilevel network meta-regression (ML-NMR). These methods use available individual patient data (IPD) to adjust for differences in effect modifiers between studies. Under much stronger assumptions, they are also used to form comparisons with single-arm studies or disconnected networks. We set out the ML-NMR framework, which extends NMA to incorporate IPD and AgD studies, and describe the advantages over previous methods. ML-NMR defines an individual-level regression model for the IPD studies and averages this model over the covariate distribution in AgD studies, avoiding aggregation bias. Unlike MAIC and STC, ML-NMR can analyse networks of any size and in larger networks allows key assumptions to be assessed. We show how ML-NMR can target either conditional or marginal estimands, depending on the needs of decision-makers. Crucially, ML-NMR can provide comparisons in any target population for decision making. This offers advantages in the context of joint clinical assessment, where a single ML-NMR analysis may be used to target the different populations required in each jurisdiction. Recent developments extend ML-NMR to incorporate single-arm studies or disconnected networks, adjusting for both effect modifiers and prognostic factors and, in larger networks, capturing additional uncertainty due to unobserved differences between studies. Approaches to incorporate additional information from trial publications such as regression coefficients or subgroup analyses have also been developed, increasing overall precision and avoiding the need for simplifying assumptions about effect modifier interactions. A user-friendly R package multinma is available for performing ML-NMR analyses.

Population-adjusted indirect comparisons in health technology assessment: Practical insights from an industry perspective

11:30-11:50

Presenter: Sarah Böhme

Authors: Sarah Böhme [1]; Robert Bauer [2]; Tobias Bluhmki [3]

Affiliations: [1] Pfizer Pharma GmbH, Berlin; [2] Novo Nordisk, Søborg; [3] Bristol Myers Squibb, München

Randomized controlled trials (RCTs) remain the gold standard for generating comparative evidence. However, indirect treatment comparisons (ITCs) are often necessary where head-to-head RCTs are unavailable. This is particularly common within health technology assessment (HTA) processes involving orphan drugs, precision medicine, rapidly evolving standards of care, or ethical and regulatory constraints. In many situations, only aggregate-level data are accessible for the comparator of interest, and differences in prognostic or effect-modifying variables across studies challenge the validity

of conventional ITC techniques. Population-adjusted indirect comparisons (PAICs) - including approaches such as matching-adjusted indirect comparison and multilevel network meta-regression - provide a toolbox to account for between-trial imbalances. Rationale, requirements and practical considerations associated with PAICs will be reflected from an industry perspective. Using practical examples, we highlight methodological assumptions, data needs, and operational challenges encountered during evidence generation within HTA timelines. Although PAICs cannot replace head-to-head RTCs, they can provide valuable insights when adequately applied. Their implementation is often restricted by limited data availability and between-trial heterogeneity, including differences in outcome definitions or study design. The introduction of the Joint Clinical Assessment under the European HTA framework adds further complexity by addressing heterogenous PICO frameworks among member states. This increases operational burden and underscores the necessity of population-adjusted methods to address a high number of PICOs requested. A persistent challenge is ensuring that all relevant prognostic and effect-modifying variables are identified, consistently measured, available, and appropriately incorporated into the analysis. Even with careful implementation, residual uncertainty remains inevitable due to the non-randomized nature of these comparisons. ITCs are important evidence for HTA submissions where direct RCT evidence is lacking. A validation of assumptions, transparent reporting, and comprehensive sensitivity analyses are essential to ensure credibility. To fully leverage the potential of PAIC methods, a concise appraisal of limitations and potential bias as well as greater alignment across HTA bodies on methodological expectations and acceptable uncertainty levels is desired. Early and iterative engagement between industry, regulatory and HTA bodies can improve planning reliability and increase the likelihood that population-adjusted indirect evidence is appropriately considered in clinical and reimbursement decision making.

Role of population-adjusted indirect comparisons in EU joint clinical assessments

11:50-12:10

Presenter/Author: David Mc Connell

Affiliations: National Centre for Pharmacoeconomics, Dublin

The EU HTA regulation introduced Joint Clinical Assessments (JCAs) of all newly authorised medicinal products in the EU, starting with oncology products and ATMPs in January 2025. The scope of the JCA is the relative effectiveness of the new product versus relevant comparators, identified by member states according to their evidence needs. For many relative effectiveness questions, direct comparative evidence will be unavailable and indirect comparisons must be considered. Population adjusted indirect comparisons (PAICs) arise naturally in this setting, particularly where individual participant (IPD) data are unavailable from comparator studies. Using examples from both national and EU HTAs, we outline the different ways in which population-adjustment may play a role in JCA. We review the relevant Methodological Guidelines of the EU HTA coordination group and summarise their implications for these methods in JCA, focusing on the distinct considerations that arise across the anchored and un-anchored cases. In the anchored setting, population-adjustment

methods are considered appropriate methodology to address effect-modification. Key challenges include appropriate pre-specification, transparency, and reliable identification of effect-modifiers. While multi-level network meta-regression has numerous advantages over other methods, its application in HTA submissions has so far been limited. Unanchored PAICs, while routinely used for indirect comparisons with single-armed trials in practice, rarely provide sufficient evidence for effect estimation due to the difficulty of ensuring that all prognostic and effect-modifying variables are adjusted for. While there is a clear role for anchored PAICs in JCA, there remains a need to improve pre-specification, variable selection, and assessment of underlying assumptions. The unanchored case is far more problematic and introduces additional challenges beyond the (already difficult) situation of traditional non-randomised studies. In all cases, access to IPD from comparator studies could improve evidence generation, and IPD-sharing initiatives should be encouraged.

AI supported Data Retrieval and Data Management in the Health Domain

Wed, 30.09.26, 11:00–12:30, Room: N3

Chair: Caroline Bönisch, Martin Golebiewski, Matthias Löbe

Authors: Martin Golebiewski [1]; Caroline Bönisch [2]; Harald Kusch [3]; Matthias Löbe [4]; Gokhan Ertaylan [5]; Wolfgang Müller [6]; Kay-Michael Otto [2]; Philip Hempel [7]

Affiliations: [1] HITS - Heidelberg Institute for Theoretical Studies, Heidelberg; [2] Hochschule Stralsund, Stralsund; [3] Universitätsmedizin Göttingen, Georg-August-Universität, Institut für Medizinische Informatik, Göttingen; [4] Universität Leipzig, Leipzig; [5] VITO - Flemish Institute for Technological Research, Mol; [6] HITS, Heidelberg, Heidelberg; [7] University Medical Center Göttingen, Göttingen

General outline and objectives

The workshop will introduce and discuss different approaches that develop workflows and tools which use Artificial Intelligence (AI), Machine Learning (ML) and Large Language Models (LLM) for supporting data retrieval, structuring, storing as well as sharing of health-related data and corresponding metadata. The aim is to showcase possibilities to integrate AI, ML and LLM-based methods in health-data management infrastructures and workflows. Attendees will be provided with insights into how such methods could be applied on what kind of data with which kind of questions. Impulse lectures will highlight the creation of an AI-generated digital twin infrastructure for healthcare in Europe and how AI methods can support the enrichment of health metadata or the usage of biosignals for clinical decision support. Also data quality aspects as well as difficulties and pitfalls of these modern technologies in the context of the health domain will be addressed. An interactive panel discussion with all speakers concludes the workshop, allowing the audience to discuss the topics with the panelists.

Structure

11:00 - 11:05 Introduction

11:05 - 11:20 Burakcan Izmirli: AI-generated digital twin infrastructure for healthcare and living environments

11:20 - 11:35 Wolfgang Müller: AI-based enrichment of health metadata: Pushing from overconfidence towards reliable outcomes

11:35 - 11:50 Kay-Michael Otto, Caroline Bönisch: The influence of data quality on the performance of AI systems

11:50 - 12:05 Philip Hempel: Explainable Artificial Intelligence on Biosignals for Clinical Decision Support

12:05 - 12:30 Panel discussion with the audience

From Building Signals to Health-Relevant Exposure Variables: An AI-Supported Retrieval Layer for Cross-Domain Digital Twins

11:05-11:20

Presenter: Burakcan Izmirli

Authors: Gokhan Ertaylan [1]; Burakcan Izmirli [2]

Affiliations: [1] VITO NV, Mol; [2] Flemish Institute for Technological Research (VITO), Belgium, Mol

The AIGENT project will explore how artificial intelligence (AI) can support data retrieval and data management across domains that are increasingly relevant for health, but that remain difficult to connect in a safe, interpretable, and standardized way. In particular, the project will address the gap between high-frequency environmental and building-related data streams, such as temperature, humidity, air quality, ventilation, occupancy, and energy operation, and health-domain data models used for care, prevention, and research. Rather than aiming to feed raw building or sensor data directly into medical decision-making systems, AIGENT will investigate an intermediate retrieval and harmonization layer based on health-relevant exposure variables. This layer is intended to translate heterogeneous technical signals into governed, semantically described, and uncertainty-aware exposure context that can be interpreted by both environmental and health-domain stakeholders. Examples include thermal comfort indicators, CO_2-derived ventilation proxies, particulate exposure summaries, temporal occupancy context, and heat-stress-related exposure profiles. The planned work will focus on three interrelated questions. First, how can AI methods support the discovery, retrieval, and contextualization of relevant environmental data from fragmented sources? Second, how can retrieved signals be mapped to standardized exposure concepts that are compatible with health data management practices? Third, how can such mappings preserve essential metadata on time, place, person- related context, consent, provenance, and uncertainty? We would like to introduce the planned AIGENT approach as a conceptual and technical contribution to AI- supported data retrieval in the health domain. It will particularly emphasize the need for shared exposure ontologies, semantic mappings, governance mechanisms, and FAIR metadata practices. As the project will not yet have generated empirical results by the time of the conference, the talk will present the project rationale, planned architecture, expected standardization challenges, and opportunities for collaboration with the dataharmonization community.

AI-based enrichment of metadata: Pushing from overconfidence towards reliable outcomes

11:20-11:35

Presenter/Author: Wolfgang Müller

Affiliations: HITS, Heidelberg, Heidelberg

The use of generative AI promises comparatively easy solutions for problems that in the past have been much too hard to tackle at all, also in the health domain. ChatGPT and related tools now enable

everyone to easily get useful data extracted from papers and online resources. Really? While generative AI looks like pure magic and perfection in the first few minutes of looking at results, verification of these initial results often will show limitations and errors. These problems arise all the more in domains that are outside the mainstream. We will give examples of our use of AI in for extracting and mapping life-science and health related data items and terminologies, and show gotchas as well as mitigation measures that permit making progress.

The influence of data quality on the performance of AI systems

11:35-11:50

Presenter/Author: Kay-Michael Otto, Caroline Bönisch

Affiliations: Hochschule Stralsund, Stralsund

Artificial Intelligence (AI) has become increasingly integrated into Health Information Systems (HIS), supporting clinical decision-making, diagnostics, predictive analytics, patient monitoring and administrative efficiency. As healthcare organizations continue to adopt AI-driven technologies, the importance of reliable and high-quality data has become more evident than ever. Although AI systems have advanced rapidly in recent years, their effectiveness, safety, and trustworthiness remain heavily dependent on the quality of the data used for training, validation, and implementation. Inaccurate, incomplete, inconsistent, or outdated data can significantly compromise AI outcomes, introduce bias, and reduce confidence in healthcare technologies among both clinicians and patients. This talk explores the critical relationship between data quality and AI performance within healthcare environments, emphasizing how poor-quality data can negatively affect clinical decision-making, operational processes, and patient outcomes. The discussion highlights key dimensions of data quality relevant to AI applications in HIS, including consistency, completeness, accuracy, timeliness, interoperability, accessibility, and semantic specificity. Particular attention is given to the risks associated with biased or unrepresentative datasets, which may reinforce existing healthcare disparities and negatively impact underrepresented patient populations. In addition, it will examine practical strategies for improving data quality and enhancing AI reliability. These strategies include data standardization, validation protocols, metadata management, governance frameworks, and continuous quality monitoring. The role of healthcare professionals, data scientists, and policymakers in maintaining ethical AI practices is also considered. Ultimately, data quality should not be viewed merely as a technical issue, but as a foundational requirement for ethical, transparent, and effective AI in healthcare. By strengthening data quality practices, healthcare institutions can improve AI system performance, minimize bias, support regulatory compliance, and contribute to safer, more equitable, and patient-centered healthcare outcomes.

Explainable Artificial Intelligence on Biosignals for Clinical Decision Support

11:50-12:05

Presenter: Philip Hempel

Authors: Philip Hempel; Dagmar Krefting

Affiliations: University Medical Center Göttingen, Göttingen

Recent advances in artificial intelligence make it possible to extract latent patterns from high-dimensional physiological signals and to relate them to cardiovascular risk, disease phenotypes, and clinically meaningful trajectories over time. However, the use of such models in clinical decision support requires more than high predictive performance. It requires transparent model behaviour, robust validation, careful consideration of signal quality, and explanations that can be compared with established physiological and clinical knowledge. This workshop focuses on explainable artificial intelligence for biosignals, with particular emphasis on electrocardiography. Current work on AI-derived ECG aging demonstrates that longitudinal ECG information can be used to capture cardiac aging effects and associate them with increased cardiovascular risk [1]. Complementary methodological developments show how explainability approaches can be adapted to biosignal models in order to support clinically interpretable decision-making [2]. Furthermore, in 12-lead ECG analysis, graph-based modelling enables the explicit representation of spatial relationships between ECG leads and supports the identification of lead-specific contributions to cardiovascular disease classification and myocardial infarction localization [3]. Further work on clinical and wearable ECG data illustrates the potential of biosignal- based AI for risk assessment in settings closer to continuous or patient-centred monitoring [4]. The workshop will discuss these developments from a methodological and translational perspective. It will address explainable AI for biosignals, the influence of data and signal quality on model performance, and the role of clinically meaningful explanations in assessing whether AI-based predictions are plausible, reproducible, and useful for decision support. The objective is to provide participants with a focused overview of how AI methods can be applied to biosignals, which assumptions and limitations must be considered, and how explainability can help bridge the gap between computational modelling and clinical interpretation. References [1] Hempel P , Ribeiro AH, Vollmer M, Bender T, Dörr M, Felix SB, et al. Explainable AI associates ECG aging effects with increased cardiovascular risk in a longitudinal population study. npj Digit Med. 20258:25. doi:10.1038/s41746-024-01428-7. [2] Maurer MC, Metsch JM, Hempel P , Bender T, Spicher N, Hauschild AC. Explainable Artificial Intelligence on Biosignals for Clinical Decision Support. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: Association for Computing Machinery 2024. p. 6597-604. doi:10.1145/3637528.3671459. [3] Maurer MC, Hempel P , Steinhaus KE, Chereda H, Vollmer M, Krefting D, et al. xGNN4MI: explainability of graph neural networks in 12-lead electrocardiography for cardiovascular disease classification. npj Digit Med. 20269:256. doi:10.1038/s41746-026-02367-1. [4] Hempel P , Riedemann GC, Ribeiro AH, Graf L, et al. Single-Lead, Single-Beat: Smartwatch ECG Risk Assessment in Decompensated Heart Failure Patients via Domain-Guided Attention. TechRxiv. 2025. doi:10.36227/techrxiv.176281047.78677029.v1.

Use of Bayesian Methods for Clinical Trial Design and Analysis (1/2)

Wed, 30.09.26, 11:00–12:30, Room: Runder Saal

Chair: Annette Kopp-Schneider

Authors: Monika Jelizarow [1]; Andrea Callegaro [2]; Björn Bornkamp [3]; Christian Stock [4]; Silvia Calderazzo [5]; James Travis [6]; Kit Roes [7, 8]; Nicky Best [9]; Andy Grieve [10]; Tim Friede [11]

Affiliations: [1] UCB Biosciences GmbH, Monheim am Rhein; [2] GSK Vaccines, Wavre; [3] Novartis Pharma AG, Basel; [4] Boehringer Ingelheim Pharma GmbH & Co. KG, Ingelheim, Germany, Ingelheim am Rhein; [5] German Cancer Research Center, Heidelberg; [6] Division of Biometrics II, Office of Biostatistics, Center for Drug Evaluation and Research, Baltimore; [7] Radboudumc, Nijmegen; [8] MA Methodology Working Party, Amsterdam; [9] GSK, UK; [10] King's College London, London; [11] Universitätsmedizin Göttingen, Göttingen

General outline and objectives

The FDA recently published a draft guidance titled "Use of Bayesian Methodology in Clinical Trials of Drug and Biological Products." This marks a major step toward broader acceptance of Bayesian methods in regulatory submissions, as the guidance is shifting towards a full acceptance of the Bayesian paradigm in the context of drug development and provides a roadmap for implementing Bayesian approaches in clinical trials. The draft guidance has attracted considerable attention and sparked discussions. Almost simultaneously, EMA issued a concept paper on the use of Bayesian methods in clinical development, indicating that attention to the topic is also shared by European regulators and is planned to develop into a future reflection paper. With the traditional paradigm potentially shifting, a discussion on the use of Bayesian methods for clinical trial design and analysis is timely.

Structure

1. Use of Bayesian methods for clinical trial design and analysis: approaches and applications

11:00-11:15 Monika Jelizarow: A pre-study look into post-study knowledge: on the use of pre-posterior distributions across clinical drug development

11:15-11:30 Andrea Callegaro: Covariate-Adjusted Dynamic Borrowing of Historical Controls in RCTs: a Vaccine Efficacy Case Study

11:30-11:45 Björn Bornkamp: Benchmarking Bayesian Shrinkage Methods for Subgroup Analysis on Clinical Trial Data

11:45-12:00 Christian Stock: Bayesian mixed models for repeated measures with informative priors: a structured workflow and software implementation

12:00-12:15 Silvia Calderazzo: Discussion

12:15-12:30 General discussion.: 2. Use of Bayesian methods for clinical trial design and analysis: current views and future perspectives

A pre-study look into post-study knowledge: on the use of pre-posterior distributions across clinical drug development

11:00-11:15

Presenter/Author: Monika Jelizarow

Affiliations: UCB Biosciences GmbH, Monheim am Rhein

Pre-posterior distributions describe our anticipated knowledge about the true treatment effect at the end of a study, before that study is conducted. They are defined conditional on the eventual study outcome (e.g. success or failure), thereby linking prior beliefs, study design, and post-study inference (Walley et al., 2015 Grieve, 2024). In this sense, they provide a way to look into future evidence (the 'post' part) before the future study is run (the 'pre' part). In this talk, I will present use cases for pre-posterior distributions across clinical drug development, with a focus on the design of proof-of-concept studies and optimisation of development plans. Through these use cases, I will clarify conceptual connections to probability of success and conditional probability of success, and share experiences with introducing pre-posterior thinking to diverse stakeholder groups. References: Walley RJ, Smith CL, Gale JD, Woodward P . Advantages of a wholly Bayesian approach to assessing efficacy in early drug development: a case study. Pharm Stat. 2015 May-Jun14(3):205-15. doi: 10.1002/pst.1675. Grieve AP . Pre-Posterior Distributions in Drug Development and Their Properties. Pharm Stat. 2024 Nov 25. doi: 10.1002/pst.2450.

Covariate Adjusted Dynamic Borrowing of Historical Controls in RCTs: a Vaccine Efficacy Case Study

11:15-11:30

Presenter/Author: Andrea Callegaro

Affiliations: GSK Vaccines, Wavre

Introduction

Background Traditional vaccine efficacy RCTs are large, costly and slow. Dynamically borrowing from historical controls can reduce enrolment, but unadjusted borrowing risks bias and inflated type I error when covariate distributions differ.

Methods

We extend the robust mixture prior (Schmidli et al., 2014) to a covariate-adjusted framework that formally links Bayesian dynamic borrowing (BDB) with causal inference. Covariate adjustment is implemented via IPTW, g-computation, or double-robust (DR) estimation before constructing the robust mixture prior, thereby preserving adaptive borrowing while mitigating covariate drift. This generalized propensity-score BDB approaches (Wang et al., 2022) and enables integration of alternative causal inference estimators.

Results

In extensive simulations and a retrospective HPV vaccine analysis, causal inference -adjusted borrowing reduced effective drift between current and historical controls, lowering bias, improving type I error control, and increasing power. DR implementations were most robust to model misspecification. Conclusions The covariate-adjusted robust mixture prior unifies BDB and causal inference, broadening reliable, practical options for leveraging historical controls in RCTs. References Schmidli H, et al. Biometrics 201470:1023-1032. Wang X, et al. J Biopharm Stat 202232:1-21.Callegaro A, et al. Pharm Stat 202423:630-644.

Benchmarking Bayesian Shrinkage Methods for Subgroup Analysis on Clinical Trial Data

11:30-11:45

Presenter/Author: Björn Bornkamp

Affiliations: Novartis Pharma AG, Basel

Subgroup analyses in clinical trials are prone to unstable estimates and findings, making shrinkage methods especially appealing. In this talk, we benchmark Bayesian shrinkage approaches for subgroup analysis using real clinical trial data. Our evaluation is based on a pair of 'twin' trials and one larger set of similar trials, where always one trial is used for learning and the other for scoring predictive performance. We compare classical hierarchical Bayesian subgroup models with global regression approaches using shrinkage priors such as the horseshoe and R2D2, following Wolbers et al. (2025). The results provide a practical comparison of how different Bayesian shrinkage methods perform in identifying and predicting subgroup effects across related clinical trials. Reference Wolbers, M., Vázquez Rabuñal, M., Li, K., Rufibach, K., & Sabanés Bové, D. (2025). Using shrinkage methods to estimate treatment effects in overlapping subgroups in randomized clinical trials with a time-to-event endpoint. Statistical Methods in Medical Research, 34(5), 903-914.

Bayesian mixed models for repeated measures with informative priors

11:45-12:00

Presenter/Author: Christian Stock

Affiliations: Boehringer Ingelheim Pharma GmbH & Co. KG, Ingelheim, Germany, Ingelheim am Rhein

The mixed model for repeated measures (MMRM) is a standard approach for analyzing continuous longitudinal endpoints in clinical trials. Despite growing interest in Bayesian borrowing in clinical trials, direct application to the MMRM has remained limited, relying, when pursued, on custom implementations. Specifying informative priors on clinically meaningful quantities in a model with treatment-by-visit interactions and additional covariates as in the MMRM is inherently difficult, and

in practice borrowing is often reduced to synthesizing evidence at a single time point in a simpler Bayesian model, leaving the longitudinal structure largely unexploited. This talk presents a workflow that addresses these challenges, covering planning, fitting, assessing, and reporting of Bayesian MMRMs with informative priors. Central to the workflow is the concept of informative prior archetypes: standardized model reparameterizations in which fixed-effect parameters correspond directly to clinically interpretable quantities, enabling deliberate and transparent prior assignment across visits and treatment arms. This correspondence reduces reduces the gap between the statistical model and domain knowledge, and facilitates prior elicitation and communication of assumptions to clinical and regulatory stakeholders. The R package brms.mmrm provides a flexible implementation of the proposed workflow, lowering the barrier to applying Bayesian MMRMs with informative priors.

13:00–14:00

Forum Digital Medicine and Artificial Intelligence (4/4)

Wed, 30.09.26, 13:00–14:00, Room: K3

Chair: Hans-Ulrich Prokosch

Authors: Hans-Ulrich Prokosch [1]; Christian Hinske [2]; Rainer Röhrig [3]; Felix Balzer [4]; Christiane Schewe [5]; Alanna Ebigbo [6]; Eimo Martens [7]; Bjoern Eskofier [8]; Daniel Röder [9]; Isabella Wiest [10]; Katharina Danhauser [11]; Hartmuth Nowak [12]; Julian Varghese [13]

Affiliations: [1] Universität Erlangen, Erlangen; [2] Lehrstuhl für Datenmanagement und Clinical Decision Support, Universität Augsburg, Augsburg; [3] Universitätsklinik RWTH Aaachen, Aachen; [4] Charité - Universitätsmedizin Berlin, Berlin; [5] Klinik und Poliklinik für Anästhesiologie, Universitätsmedizin Rostock, Rostock; [6] Klinik für Innere Medizin, Katholisches Klinikum Bochum, Bochum; [7] Klinik und Poliklinik für Innere Medizin I, University Hospital, Technical University of Munich, TUM School of Medicine and Health, Munich, Munich; [8] Lehrstuhl für KI-unterstützte Therapieentscheidungen, Ludwig-Maximilians-Universität Münch, München; [9] Anästhesiologischen Intensivstation, Universitätsklinikum Würzburg, Würzburg; [10] Else Kröner Fresenius Zentrum für Digitale Gesundheit TU Dresden, Dresden; [11] LMU Munich, München; [12] Universitätsklinikum Knappschaftskrankenhaus Bochum, Bochum; [13] Institute of Medical Data Science, Otto von Guericke University Magdeburg, Magdeburg

General outline and objectives

The workshop is part of the Forum Digital Medicine and Artificial Intelligence, which aims to bring together Clinformaticians (Clinical Informaticians) from the clinical societies of the AWMF with the methodological disciplines of the GMDS in order to discuss interdisciplinary methodological, regulatory, and political challenges in digital medicine as well as for the development and application of Artifical Intelligence Tools. Based on stimulating expert presentations the workshop participants will discuss the current challenges and hurdles and interactively develop solutions for future developments and activities. One part of the workshop will be a Pro&Contra Presentation and Discussion on the opportunities, but also the challenges and hurdles for agentic AI systems.

Structure

12:00 - 12.30 Daniel Röder,

DGAI WAK Digital Medicine and data-driven telemedicine

13:00 - 14:00 Workshop Pro&Contra

13:00 - 13:30 Christopher Sauer: Pro Position in the Pro&Contra Session on Agentic AI Applications in Healthcare

13:30 - 14:00 Jürgen Schulze: Contra Position in the Pro&Contra Session on Agentic AI Applications in Healthcare

Pro Position in the Pro&Contra Session on Agentic AI Applications in Healthcare

13:00-13:30

Presenter/Author: Christopher Sauer

Affiliations: Institut für künstliche Intelligenz in der Medizin (IKIM), Universitätsklinikum Essen, Universität Duisburg-Essen, Essen

KI-Agenten sind (semi-)autonome Systeme, die komplexe mehrstufige Aufgaben mit minimalem menschlichen Eingriff bewältigen und ganze klinische Workflows orchestrieren: von der Auswertung von Patientenakten über die Synthese medizinischer Leitlinien bis hin zur Vorbereitung klinischer Entscheidungen. Erste Proof-of-concept-Studien belegen ihr Potenzial, etwa in der personalisierten Therapieplanung, der automatisierten Zuordnung zu klinischen Studien oder der spürbaren Entlastung ärztlicher Routinearbeit. Die Frage ist nicht mehr ob, sondern wie wir diese Transformation gestalten. Diese Präsentation liefert Argumente für den Einsatz von agentischer KI.

Contra Position in the Pro&Contra Session on Agentic AI Applications in Healthcare

13:30-14:00

Presenter/Author: Jürgen Schulze

Affiliations: German University of Digital Science, Babelsberg

"Enjoying the good in the bad while understanding the bad in the good' (Das Gute im Schlechten genießen und gleichzeitig das Schlechte im Guten begreifen) Nur wenige disruptive Technologien haben es schneller in Breite und Tiefe und damit in die, zumindest gefühlte, Unverzichtbarkeit geschafft, wie GenAI und dessen Posterchild ChatGPT. Bisher automatisierte diese Technologie bis zu einem gewissen Grade Analyse, Entscheidungsfindung und Kreation - im Bildlichen wie im Textlichen. Agenten legen nach: Sie setzen um. Ob die Gestaltung der verwendeten Architekturen den Anforderungen an Befähigung der Anwendenden und ethischem sowie technisch/legalem Regelwerk gerecht wird, geht in der aktuellen shiny-Tech-Euphorie unter. Wo jetzt ein pragmatischer Ansatz hilfreich wäre, herrscht derzeit ein Klima der Extrempositionen durch alle Bevölkerungsschichten und Expertendisziplinen hindurch. Dabei wird auf allen Seiten mit Gefühlen hantiert, dem Offensichtlichen und Unmittelbaren, und damit der Diskursraum für die nicht gestellten Fragen verstopft. Wir müssen wieder lernen zu differenzieren, Interdependenzen zu beleuchten und gedanklich über Bande zu spielen, um Lösungen zu finden, mit den Risiken umzugehen, ohne auf die Segnungen der Technologie zu verzichten.

14:00–15:30

Forum Digital Medicine and Artificial Intelligence (3/4)

Wed, 30.09.26, 14:00–15:30, Room: K3

Chair: Hans-Ulrich Prokosch

Authors: Hans-Ulrich Prokosch [1]; Christian Hinske [2]; Rainer Röhrig [3]; Felix Balzer [4]; Christiane Schewe [5]; Alanna Ebigbo [6]; Eimo Martens [7]; Bjoern Eskofier [8]; Daniel Röder [9]; Isabella Wiest [10]; Katharina Danhauser [11]; Hartmuth Nowak [12]; Julian Varghese [13]

Affiliations: [1] Universität Erlangen, Erlangen; [2] Lehrstuhl für Datenmanagement und Clinical Decision Support, Universität Augsburg, Augsburg; [3] Universitätsklinik RWTH Aaachen, Aachen; [4] Charité - Universitätsmedizin Berlin, Berlin; [5] Klinik und Poliklinik für Anästhesiologie, Universitätsmedizin Rostock, Rostock; [6] Klinik für Innere Medizin, Katholisches Klinikum Bochum, Bochum; [7] Klinik und Poliklinik für Innere Medizin I, University Hospital, Technical University of Munich, TUM School of Medicine and Health, Munich, Munich; [8] Lehrstuhl für KI-unterstützte Therapieentscheidungen, Ludwig-Maximilians-Universität Münch, München; [9] Anästhesiologischen Intensivstation, Universitätsklinikum Würzburg, Würzburg; [10] Else Kröner Fresenius Zentrum für Digitale Gesundheit TU Dresden, Dresden; [11] LMU Munich, München; [12] Universitätsklinikum Knappschaftskrankenhaus Bochum, Bochum; [13] Institute of Medical Data Science, Otto von Guericke University Magdeburg, Magdeburg

General outline and objectives

The workshop is part of the Forum Digital Medicine and Artificial Intelligence, which aims to bring together Clinformaticians (Clinical Informaticians) from the clinical societies of the AWMF with the methodological disciplines of the GMDS in order to discuss interdisciplinary methodological, regulatory, and political challenges in digital medicine as well as for the development and application of Artifical Intelligence Tools. Based on stimulating expert presentations the workshop participants will discuss the current challenges and hurdles and interactively develop solutions for future developments and activities. One part of the workshop will be a Pro&Contra Presentation and Discussion on the opportunities, but also the challenges and hurdles for agentic AI systems.

Structure

14:00 - 15:30 Workshop-Part 3

14:00 - 14:30 Katharina Danhauser: Digital Assistance in the Doctor's Office: How LLMs Support Doctors

14:30 - 15:00 Hartmuth Nowak: Generating clinical value with AI

15:00 - 15:30 Julian Varghese: AI-Competencies in Medical Education - The Role of Medical Data Science

Digital Assistance in the Doctor's Office: How LLMs Support Doctors

14:00-14:30

Presenter/Author: Katharina Danhauser

Affiliations: LMU Munich, München

Large Language Modelle können Ärztinnen und Ärzte bei administrativen sowie klinischen Aufgaben unterstützen. Das mögliche Anwendungsspektrum umfasst eine automatisierte Dokumentation und Erstellung von Arztbriefen sowie die Unterstützung bei der Informationsrecherche und Patientenkommunikation. Hierdurch soll eine eine Optimierung der Arbeitsabläufe und eine Entlastung des medizinischen Personals erzielt werden, wodurch mehr Zeit für die direkte Patientenversorgung zur Verfügung steht. Gleichzeitig bestehen Herausforderungen hinsichtlich des Datenschutzes, der Transparenz, der medizinischen Validität und möglicher fehlerhafter Inhalte. Der Beitrag zeigt Chancen und Risiken des Einsatzes von LLMs im klinischen Alltag. Large language models have the potential to support physicians in administrative and clinical tasks. Potential applications include automated documentation and the generation of medical reports, as well as support with information retrieval and patient communication. The objective is to enhance workflows and alleviate the workload on medical staff, thereby allocating more time to direct patient care. However, there are challenges regarding data protection, transparency, medical validity, and potentially erroneous content. This contribution provides opportunities and risks associated with the integration of LLMs into routine clinical practice.

Generating clinical value with AI

14:30-15:00

Presenter/Author: Hartmuth Nowak

Affiliations: Ruhr University Bochum, Knappschaft Kliniken University Hospital Bochum, Department of Anesthesiology, Intensive Care Medicine and Pain Therapy, Center for Artificial Intelligence, Medical Informatics, Bochum

Artificial intelligence has become a driving force in discussions about the future of healthcare, yet enthusiasm for the technology itself must not obscure the fundamental question: what tangible value does it actually deliver? AI is not an end in itself, but a means to generate measurable clinical and, where applicable, economic benefit. The central criterion for any digital intervention - whether AI-powered or not - must therefore be the value it creates for patients, clinicians, and healthcare systems alike. Complexity should never be introduced for its own sake. If a simpler, rule-based or statistical approach achieves the same or even superior outcomes compared to a sophisticated machine learning model, the simpler solution is preferable. Unnecessary complexity increases implementation risk, reduces transparency, and impedes clinical adoption. The guiding principle must always be fitness for purpose, not technical ambition. Three concrete use cases illustrate how value-driven AI implementation can work in practice. Early sepsis detection demonstrates how AI-based alert

systems can significantly reduce time-to-treatment and improve patient outcomes. Digital twins for patients with complex or chronic conditions enable individualized simulations of disease trajectories and treatment responses, supporting more precise therapeutic decisions. Clinical decision support systems (CDSS) for early complication detection can assist clinical teams in identifying deteriorating patients before critical events occur, adding a further dimension of proactive, data-driven care. Across all three domains, a recurring challenge emerges: demonstrating value in isolated settings is not sufficient. For AI-based interventions to become part of standard care, they must find their way into clinical guidelines. This can only be achieved through systematic evidence generation. Prospective clinical evaluations that measure not only technical performance but real-world patient outcomes are indispensable. Regulatory frameworks and guideline bodies are increasingly receptive to such evidence, but the burden of proof lies with those developing and deploying these systems. Ultimately, a value-first mindset is needed in clinical AI: rigorous evaluation, honest benchmarking, and a clear commitment to putting clinical benefit before technological novelty.

AI-Competencies in Medical Education - The Role of Medical Data Science

15:00-15:30

Presenter/Author: Julian Varghese

Affiliations: Institute of Medical Data Science, Otto von Guericke University Magdeburg, Magdeburg

(german version below) Medical data science is an integral part of medical school curricula. It provides future physicians with a solid foundation in the areas of digital health literacy and artificial intelligence. In this context, new, tailored learning objectives are presented that address recent developments such as large language models and generative AI. I will also outline the extent to which and the level of proficiency at which these topics can be taught in the classroom. Die medizinischen Datenwissenschaften sind ein fester Bestandteil des Medizinstudiums. Sie liefern angehenden Ärztinnen und Ärzten wichtige Grundlagen in den Bereichen Digital Health Literacy und Künstliche Intelligenz. In diesem Kontext werden neue, abgestimmte Lernziele präsentiert, die jüngste Entwicklungen wie Large Language Models oder Generative KI aufgreifen. Zudem wird dargestellt, in welchem Umfang und auf welchem Kompetenzlevel diese Themen im Unterricht vermittelt werden können.

What Makes a Trial Complex? Evidence, Experience, and AI

Wed, 30.09.26, 14:00–15:30, Room: K4

Chair: Franz Koenig, Frank Pétavy, Tim Friede

Authors: Franz Koenig [1]; Fabian Eibensteiner [1]; Frank Pétavy [2]; Anouk Neven [2]; Tim Friede [3]; Florian Klingmueller [4]; Sonja Drescher [3]; Thomas Jaki [5]; Martin Posch [1]; Martin Hellmich [6]

Affiliations: [1] Medical University of Vienna, Vienna; [2] European Medicines Agency, Amsterdam; [3] University Medical Center Göttingen, Göttingen; [4] Austrian Medicines and Medical Devices Agency (MEA / AGES), Vienna; [5] University of Regensburg, Regensburg; [6] Department of Medical Statistics, University Medical Center Göttingen, Göttingen

General outline and objectives

This workshop explores clinical trial complexity as a dynamic and evolving concept rather than a fixed classification. It examines multiple dimensions of complexity, including ethical, regulatory, statistical, methodological, and operational aspects. These are informed by findings from (a) a structured literature review, (b) a multi-stakeholder Delphi process, and (c) a review of relevant regulatory guidance. A central objective is to critically discuss whether criteria exist that make one trial inherently more complex than another, or whether "complexity" often acts as a surrogate for limited experience, emerging methodology, or infrequent use of specific designs or technologies. Participants will also learn how AI methods, particularly Large Language Models (LLMs), can support the automated identification and characterization of potentially complex trials registered in EU via the Clinical Trials Information System (CTIS, https://euclinicaltrials.eu/), while understanding the limitations and governance challenges of such approaches.

Structure

Introduction: What does "complex" mean?

Evidence from published scientific literature: How is complexity defined and operationalized?

Regulatory guidance perspectives on complex and innovative trials

Conceptual discussion: Are there objective criteria for complexity, or is it rather experience-driven?

AI and LLMs in practice: Automated identification of complex trials in CTIS

Case studies illustrating different dimensions of complexity

Live audience survey and discussion

Panel discussion and synthesis of key messages

Complexity: an opportunity for regulators?

14:00-14:15

Presenter: Frank Pétavy

Authors: Frank Pétavy; Anouk Neven

Affiliations: European Medicines Agency, Amsterdam

Complex clinical trials are frequently seen as an opportunity to expedite drug development but they are difficult to apprehend. This so-called complexity covers a range of areas, from ethics, clinical trial design (or methodology), operations, statistical analysis, and regulatory assessment and decision-making. Currently, there exists no clear definition or system for the characterisation of complex clinical trials in the regulatory context. It applies to scientific advice or evaluation of a marketing authorisation request, but also to clinical trial approval, where there are at least two ways to submit a complex clinical trial in the Clinical Trial Information System as one trial or separate trials. This talk will highlight opportunities to better understand what is complex and who can benefit from it.

Published Evidence guiding the Characterization and Classification of Complex Clinical Trials - Results from a Literature Review

14:15-14:30

Presenter: Martin Hellmich

Authors: Martin Hellmich [1]; Sonja Drescher [2]; Long-Hao Xu [2]; Tim Friede [3]

Affiliations: [1] Department of Medical Statistics, University Medical Center Göttingen, Göttingen; [2] University Medical Center Göttingen, Göttingen; [3] Universitätsmedizin Göttingen, Göttingen

This presentation summarizes results from the literature review within the "Characterisation of Complex clinical trials in the European regulatory context" project.

This review systematically evaluates how trial complexity is described and conceptualized in the scientific literature to support an evidence-informed characterization framework. The literature review was conducted by searching the PubMed database, complemented by AI-assisted semantic searches across five predefined dimensions of complexity (methodological, statistical, operational, regulatory, and ethical) using the Elicit platform. Following independent double-screening of titles and abstracts, eligible full-text articles were evaluated using a structured guide to extract verbatim text, subdimensions, and short thematic descriptions. A large language model assisted in merging the thematic clusters into cohesive descriptions.

The talk will focus on how complex clinical trials are conceptualized and described in scientific literature, highlighting their multi-dimensional nature across methodological, operational, regulatory, statistical, and ethical domains.

What Makes a Trial Complex? Mapping Trial Complexity in Regulatory Guidance

14:30-14:45

Presenter: Florian Klingmueller

Authors: Florian Klingmueller [1]; Susanne Urach [1]; Tobias Fellinger [1]; Andrew C. Hooker [2]

Affiliations: [1] Austrian Medicines and Medical Devices Agency (MEA / AGES), Vienna; [2] Uppsala University, Uppsala, Sweden

This presentation summarizes results from the regulatory guidance review within the "Characterisation of Complex clinical trials in the European regulatory context" project. It highlights the main terms and concept categories identified across key complexity dimensions, including design, statistical, operational, regulatory and ethical aspects of trial complexity. We present how individual regulatory documents map against these dimensions and what this reveals about the regulatory framing of complex clinical trials.

The talk will focus on how these results can support a more structured classification and categorization of complex trials, including the development of candidate attributes, terminology frameworks and future CTIS-relevant concepts.

What makes a trial complex? From Delphi consensus to AI-supported CTIS identification.

14:45-15:05

Presenter: Franz Koenig

Authors: Franz Koenig; Fabian Eibensteiner

Affiliations: Medical University of Vienna, Vienna

This presentation connects two steps in the characterisation of complex clinical trials (CCT): reaching expert consensus on what makes a trial complex, and applying that framework to identify complex trials in the EU Clinical Trials Information System (CTIS, see https://euclinicaltrials.eu/). First, it summarises findings from a multi-stakeholder Delphi process, which identified key dimensions of complexity, including design, statistical, operational, regulatory, and ethical considerations. The presentation then shows how these Delphi-derived features were used in manual and AI-supported review of CTIS study protocols. It will discuss how Large Language Models can support the identification of potentially complex trials, while also highlighting practical challenges such as document availability, variable terminology, ambiguous cases and the need for human oversight. Finally, the talk will briefly outline how these findings can inform future CTIS visibility, for example through structured CCT information and prioritised descriptors.

Teaching Statistical Literacy in Times of Political Changes

Wed, 30.09.26, 14:00–15:30, Room: K5

Chair: Ursula Berger, Carolin Herrmann

Authors: Ursula Berger [1]; Carolin Herrmann [2]; Walter J. Radermacher [3]; Katja Ickstadt [4]; Sabine Hoffmann [3]; Geraldine Rauch [5]; Ernst C. Wit [6]

Affiliations: [1] LMU München, München; [2] Heinrich-Heine-Universität Düsseldorf, Düsseldorf; [3] Department of Statistics, Ludwig Maximilian University of Munich (LMU Munich), Germany, München; [4] TU Dortmund University, Dortmund; [5] Charité Universitätsmedizin Berlin, Berlin; [6] Università della Svizzera Italiana Lugano, Lugano

General outline and objectives

In times, where artificial intelligence (AI) and data-driven research reshape our views, we see political dynamics that threaten the availability and credibility of statistical data sources. The value of statistical integrity and accuracy is getting increasingly important and with it the need for teaching statistics literacy to a broad audience. This session addresses current challenges and innovations in teaching statistical literacy in times of rapid AI development and increasing political challenges, with an emphasis on life sciences.

Among other topics, we want to discuss how teaching methods and curricula must adapt, the role of ethics and responsible research practices in statistical education, how civic data literacy can be strengthened, and the role of data in democracy.

Following five inspirational talks by leading representatives from statistics and data science (10 min each), a panel discussion will provide a platform to exchange perspective and discuss strategies that prepare both students and the broader public for critical engagement with technological innovations and the societal challenges surrounding data, enabling well informed and evidence-based decision-making.

Structure

14:00 - 14:10 Walter J. Radermacher: Data ethics: the missing puzzle piece in applied statistics education

14:10 - 14:20 Katja Ickstadt: Rethinking statistics education at universities

14:20 - 14:30 Sabine Hoffmann, Anne-Laure Boulesteix: Teaching responsible research practices that improve the communication and understanding of evidence and uncertainty

14:30 - 14:40 Geraldine Rauch: Democracy under attack - Empirical research and statistical inference in times of fake news

14:40 - 14:50 Ernst C. Wit: Nature of Statistical Evidence: Lessons from Covid-19

14:50 - 15:30 Panel discussion

Data ethics: the missing puzzle piece in applied statistics education

14:00-14:10

Presenter/Author: Walter J. Radermacher

Affiliations: Department of Statistics, Ludwig Maximilian University of Munich (LMU Munich), Germany, München

In an uncertain new world, attitudes towards quantitative expertise have become polarised. The credibility of statistics is no longer solely determined by the rigour of its methods it is also shaped by how clearly statisticians communicate what they measure, why they measure it, and how they transform this into trustworthiness. From one perspective, basing politics on statistics is elitist and undemocratic because it disregards people's emotional connections to their community. However, from the other perspective, statistics are quite the opposite of elitist. They enable journalists, citizens, entrepreneurs, and politicians to discuss society as a whole in a way that can be validated rather than based on anecdote, sentiment, or prejudice. Quantification has become the dominant language of governance, finance, public health, education and environmental policy. However, this proliferation has been accompanied by growing concerns about statistical malpractice, algorithmic bias, and the unintended consequences of metrics. There is also a tendency for quantification to transform political issues into technical ones by outsourcing decision-making authority to supposedly neutral, number-based systems. Sheila Jasanoff's distinction between 'technologies of hubris' and 'technologies of humility' highlights this issue: hubristic quantification promises management and control, even in conditions of deep uncertainty. This can lead to overconfidence, prevent political discussion and mask the assumptions embedded in its framing. In contrast, humble quantification openly reflects on its limitations, identifies winners and losers, and creates conditions for genuine societal learning. The authority granted by quantification means that the methodological choices involved in designing statistics are never neutral. Rather, they allocate attention, recognition and legitimacy. These considerations necessitate a transition from a sociology of quantification to an ethics of numbers. Quantification is a fundamentally social activity with ethical implications that must be explicitly and transparently addressed. Such a reflective perspective on the profession should also be incorporated into relevant educational programmes, which is not currently the case. Therefore, we all face the shared challenge of changing this situation, which cannot be addressed merely through token ethics courses in the curriculum. Instead, a joint effort is needed involving professionals from disciplines such as sociology and ethics, as well as cooperation between universities.

Rethinking statistics education at universities

14:10-14:20

Presenter/Author: Katja Ickstadt

Affiliations: TU Dortmund University, Dortmund

In the age of Big Data, Data Science, Machine Learning and Artificial Intelligence, statistics education faces new challenges in a wide range of subjects. This has led to reform efforts worldwide with the aim of establishing statistical literacy as an indispensable core competence. In Germany, many statistics courses at university level still follow a rather classical curriculum, focusing on probability theory and the application of statistical methods. Moreover, statistics is often taught within application areas such as economics or medicine, which hampers the development of generally applicable curricula. To provide guidance to statistics teachers at German universities, the Deutsche Arbeitsgemeinschaft Statistik (DAGStat), an umbrella organization for 16 statistical societies and professional associations founded a working group, which proposes the following key points to be considered in statistics education (see Berger, U. et al. 2025): Data and statistical literacy are core competencies in our data-driven world. They are interdisciplinary, and should be taught across all educational levels. Statistical analyses and the interpretation of results need to be communicated transparently and appropriately for the target audience. Moreover, a critical approach to data and data analysis should be practiced using real-world examples, and the importance of data quality needs to be stressed. Also, ethical aspects in data quality and analysis need to be considered and the resulting responsibilities communicated. Last but not least the mastery of modern software plays a central role in statistics education. One focus of the data and statistical literacy education at TU Dortmund University is on interdisciplinary tandem teaching in form of a seminar or case-based project work. Topics comprise text mining, algorithmic accountability or visualisation of sports data. The goal is to deepen both technical understanding and critical reflection, fostering a more specialized and application-oriented competence. Materials are developed as Open Educational Resources. In conclusion, we advocate to teach statistical literacy across subjects and to all levels of education in order to promote a data literate society. Reference: Berger, U. et al. (2025): Daten, Künstliche Intelligenz und Evidenz - neue Anforderungen an die Statistikausbildung an Hochschulen, AStA Wirtschafts- und Sozialstatistisches Archiv, to appear.

Teaching responsible research practices that improve the communication and understanding of evidence and uncertainty

14:20-14:30

Presenter: Sabine Hoffmann

Authors: Anne-Laure Boulesteix

Affiliations: LMU Munich, Munich

One of the fundamental roles of science is to act as a neutral and independent source of information. To address the challenges and crises of the 21st century, scientific evidence needs to be presented in a way that is accessible and understandable. At the same time, many statistical quantities that are commonly used are inherently difficult to interpret. Misinterpretations of statistical significance and of classical effect estimates, including odds ratios and hazard ratios, are so widespread that applied researchers are sometimes startled when they hear the correct definition of these quantities. If the communication of scientific evidence is difficult, the communication of scientific uncertainty is

even more intricate. Instead of accounting for the uncertainties in their data, biomedical researchers commonly ignore these uncertainties and qualitatively discuss the biases arising from the fact that uncertainties are ignored. This practice makes scientific uncertainty essentially inaccessible to readers who have limited knowledge and understanding of methodological weaknesses and challenges. Adding to this, throughout their education and when reading the scientific literature, students may get the impression that there is a unique and correct analysis strategy for every data analysis task and that this analysis strategy will always yield a significant and noteworthy result. As a consequence, they may often be ill-equipped for real-word data analysis tasks in which data are messy and there is a multiplicity of possible analysis strategies. This multiplicity of possible analysis strategies will lead to overoptimistic and nonreplicable research findings if it is combined with result-dependent selective reporting. Finally, there has been great progress in recent years to improve the transparency, reproducibility, and rigour of confirmatory works through pre-registrations, registered reports, replication and many-analyst studies. However, these solutions have left unaddressed a crucial gap: most research is simply not mature enough for confirmatory testing. This talk will reflect on ways in which we can teach statistics and statistical literacy that will improve the communication and understanding of scientific evidence and uncertainty to improve evidence-based decision making.

Democracy under attack -Empirical research and statistical inference in times of fake news

14:30-14:40

Presenter/Author: Geraldine Rauch

Affiliations: Charité Universitätsmedizin Berlin, Berlin

The trust in academia and research is decreasing worldwide. Fake news is spreading, in particular over social media. Polarization and populism in politics and media increase, for some areas such as research on climate change, research on the corona pandemic or research on migration and integration, the individual opinion formation seems no longer based on empirical research results but on popular politic viewpoints spread by the media. However, in a world flooded with fake news, the role of empirical research results and objective statistical analyses becomes more and more important. Statistician must not only provide valid analyses, they are also part of a research team which is responsible to present the empirical research results to the public in a way that the society can work with the information. On the other hand, it is also the role of statistician to highlight when facts or results are wrong or presented in a misleading context. Importantly, the role of statisticians is thus no longer restricted to a research bubble, where a statistical service is provided to for example to medical researchers. Explaining empirical results is a much broader societal need. We discuss examples how statisticians and empirical researchers can contribute to a vivid knowledge exchange with society thereby helping to stabilize objectivity as a fundamental concept of our democracy.

Nature of Statistical Evidence: Lessons from Covid-19

14:40-14:50

Presenter/Author: Ernst C. Wit

Affiliations: Università della Svizzera Italiana Lugano, Lugano

This presentation examines the nature of statistical evidence, focusing on how evidence interacts with background beliefs, media dynamics, and crowd thinking, and how this should affect the teaching of statistical ideas. Drawing on the paper "Some statistical aspects of the Covid-19 response", the talk argues that statistical evidence tends not to be interpreted in isolation, but through psychological and social filters that strongly shape public perception and policy decisions. Using examples from the UK Covid-19 response, we discuss cases where statistically questionable analyses or presentations appear to have had substantial influence on government policy, public behaviour, or both. Particular attention is given to the asymmetric communication of risk. Covid-related risks were frequently presented through worst-case projections, cumulative indicators, and highly salient statistics, while the risks associated with interventions - including educational, economic, social, and non-Covid health consequences - were often less visible or insufficiently quantified. The presentation revisits several statistical issues surrounding excess mortality, life-loss calculations, case-based indicators, and epidemic modelling. We discuss how limited attention to heterogeneity, uncertainty, and behavioural adaptation may have affected both model interpretation and policy evaluation. Relatedly, the talk addresses a central question for teaching of (bio)statistics: although it is important to stress the importance of common sense in evaluating the quality of statistical evidence, at the same time we should avoid common sense blind spots. We argue that evidence-based decision making requires not only statistical rigour, but also critical reflection on how evidence is framed, communicated, and socially received. Any statistician should to be taught to be aware of this. The Covid-19 experience illustrates that the impact of statistical evidence depends not only on the quality of the analysis itself, but also on the cognitive and collective environments in which that evidence circulates.

Deep Dive Into eHealth With the Project MINT Meets Medicine

Wed, 30.09.26, 14:00–15:30, Room: N3

Chair: Madeline Bartsch

Authors: Mattea Müller [1]; Michael Marschollek [1]; Madeline Bartsch [2]

Affiliations: [1] Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Hannover Medical School, Hannover, Germany., Hannover; [2] Peter L. Reichertz Institut für Medizinische Informatik der TU Braunschweig und der Medizinischen Hochschule Hannover, Hannover

General outline and objectives

This workshop presents results from the MINT Meets Medicine Deep Dive Challenges, an interdisciplinary training initiative at the interface of STEM disciplines and medicine hosted at the Peter L. Reichertz Institute for Medical Informatics, Hannover. Within the program, six interdisciplinary groups of early-career researchers collaborated over several months on three clinically motivated eHealth data challenges. The first challenge analyzed continuous glucose monitoring data to investigate individual glucose responses in daily life. The second explored early warning systems for chronically ill patients, using large-scale, longitudinal, synthetic electronic health record data to identify risk patterns and support proactive care. The third integrated multiomics data from kidney disease patients, including transcriptomics, to study molecular mechanisms and cell-type-specific signatures.

The workshop presents the Deep Dive Challenges results and discusses the research approaches. It will be held in a modified World Café format. Each of the six project teams hosts a discussion table where participants can engage with the team's data, research question, analytical approach, results, and open challenges.

The workshop is designed for ISCB GMDS participants from medical informatics, biostatistics, epidemiology, computer science, clinical research, and related fields. It highlights how heterogeneous health data can be analyzed in interdisciplinary teams and translated into clinically meaningful insights.

Structure

14:00 - 14:15 Madeline Bartsch: Introduction to MINT Meets Medicine and overview of the Deep Dive Challenges

14:15 - 15:15 Six interdisciplinary Deep Dive Challenge teams: Interactive World Café: Results from the MINT Meets Medicine Deep Dive Challenges

15:15 - 15:25 Table rapporteurs / M3 organizing team: Plenary synthesis: Key findings, open questions, and translational implications

15:25 - 15:30 M3 organizing team: M3 Prize and closing remarks

---

Continuous Glucose Monitoring in Diabetes Mellitus: Insights from Interdisciplinary Analysis of Free-Living Multimodal Data

14:15-15:15

Presenter/Author: Sophia Schulmeister, Hannah Brüx, Sinan Yilmaz, Jiska Ziegler, Jan Sören Breidenbach, Ann-Kristin Zoeller, Lene Bleich

Continuous glucose monitoring (CGM) enables detailed characterization of glycemic dynamics in real-world settings, yet translating these complex time-series data into clinically meaningful insights remains challenging. Within the MINT Meets Medicine (M3) Deep Dive Challenges, two interdisciplinary groups of early-career researchers have collaboratively analyzed CGM data to investigate determinants of individual glucose responses in daily life. The challenge was based on the CGMacros dataset (1), a multimodal free-living dataset integrating continuous glucose measurements with detailed information on meal composition (macronutrients), physical activity, demographic variables, and clinical characteristics. The dataset includes healthy individuals, subjects with prediabetes, and patients with type 2 diabetes, enabling comparative analyses across metabolic groups. The primary objective was to assess how meal composition, time of day, and physical activity influence postprandial glucose responses and to identify patterns distinguishing different metabolic states. Two different groups applied a range of analytical approaches, including exploratory time-series analysis, visualization of glycemic metrics, clustering of postprandial glucose trajectories, and predictive modeling of glucose excursions using statistical and machine learning methods. In addition to methodological exploration, the challenge emphasized integrating heterogeneous data sources, handling temporal dependencies, and translating analytical findings into clinically meaningful interpretations. This challenge aims to demonstrate the potential of combining CGM with lifestyle data for personalized metabolic assessment and highlights the value of interdisciplinary collaboration in bridging data science methods and clinical reasoning in eHealth research. 1. Das A, Kerr D, Glantz N, Bevier W, Santiago R, Gutierrez-Osuna R, et al. CGMacros: a pilot scientific dataset for personalized nutrition and diet monitoring. Sci Data. 2025 Sep 2512(1):1557. doi:10.1038/s41597-025-05851-7

Multiomics Analysis in Kidney Disease: An Interdisciplinary Data Challenge Integrating Bulk, Single-Cell, and Spatial Transcriptomics

14:15-15:15

Presenter/Author: Jan Weber, Nicolas Hebestreit, Nikolas Kühn, Jan Tauwaldt, Pedro Octavio Markus Gebhard Apalategui, Bendix Harms, Julius Hellwig, Tristan Baumann

Understanding the biological mechanisms underlying heterogeneous disease trajectories in kidney disease remains a central challenge in precision medicine. Advances in transcriptomics now enable the investigation of disease processes at multiple levels, from global gene expression patterns to cell-type-specific and spatially resolved molecular signatures. Within the MINT Meets Medicine (M3) Deep Dive Challenges, two interdisciplinary groups of early-career researchers will explore multiomics approaches to investigate kidney disease mechanisms. The challenge was based on data from

the Kidney Precision Medicine Project (KPMP), a comprehensive, publicly available resource integrating bulk RNA sequencing, single-cell and single-nucleus transcriptomics, and spatial transcriptomics, complemented by clinical and histopathological annotations (1). This multi-layered dataset provides a unique opportunity to analyze disease processes across different biological scales. The primary objective was to investigate how transcriptomic data can contribute to a better understanding of disease mechanisms, cell-type-specific alterations, and potential targets for diagnosis or therapy in kidney disease. Therefore, two groups performed exploratory analyses of gene expression patterns, identified differences between disease and reference samples, and investigated cell-type-specific or region-specific signatures. More advanced approaches may include integrating multiple data modalities, such as linking bulk and single-cell data or combining cell-type information with spatial context, to generate hypotheses about the cellular and molecular drivers of disease. Emphasis was placed on the biological interpretation of findings and their connection to clinically relevant questions. This challenge highlights the potential of multiomics data to advance precision medicine and underscores the importance of interdisciplinary collaboration in translating complex molecular data into clinically meaningful insights. 1. Hansen J, Sealfon R, Menon R, Eadon MT, Lake BB, Steck B, et al. A reference tissue atlas for the human kidney. Sci Adv. 2022 Jun 108(23):eabn4965. doi:10.1126/sciadv.abn4965

Early Warning Systems for Chronic Disease Management: An Interdisciplinary Data Challenge Using Synthetic Electronic Health Records

14:15-15:15

Presenter/Author: Elisheva Naor, Vera Schwane, Adrian Bardtke, Joshua Georg Tellbach, Delight Kurian-Chandy, Sorelle Fekam Kammgne, Sophie Pernickel, Louisa Eskelson, Martin Eberhardt, Luca Busch

The increasing availability of electronic health records (EHRs) offers significant potential to improve the management of chronic diseases, particularly by enabling early identification of patients at risk of deterioration. However, the complexity and heterogeneity of longitudinal health data present substantial analytical and interpretative challenges. Within the MINT Meets Medicine (M3) Deep Dive Challenges, two interdisciplinary groups of early-career researchers have explored approaches for developing data-driven early warning systems using large-scale synthetic EHR data. The challenge was based on data generated with Synthea, an open-source framework that simulates comprehensive, longitudinal patient records, including diagnoses, procedures, medications, laboratory values, healthcare encounters, and demographic and social determinants (1). These data provide a realistic yet privacy-preserving environment for methodological exploration. The primary objective was to investigate how large-scale EHR data can be used to identify patients with chronic diseases who are at increased risk of complications or adverse outcomes. Two research groups focused on clinically relevant conditions (e.g., diabetes mellitus or chronic heart failure) and performed exploratory analyses of patient cohorts, investigate longitudinal disease trajectories, and identify patterns associated with disease progression. Building on these analyses, the groups developed and compared approaches for early risk detection, ranging from interpretable statistical models to machine learning-based prediction frameworks. Particular emphasis was placed on integrating heterogeneous data sources,

temporal modeling of patient histories, and the clinical interpretability of predictive approaches. This challenge illustrate the opportunities and limitations of using large-scale health data for proactive care and highlights the importance of interdisciplinary collaboration in translating data-driven models into clinically meaningful decision support in eHealth. 1. Walonoski J, Kramer M, Nichols J, Quina A, Moesel C, Hall D, et al. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. J Am Med Inform Assoc. 2018 Mar 125(3):230-8. doi:10.1093/jamia/ocx079

Use of Bayesian Methods for Clinical Trial Design and Analysis (2/2)

Wed, 30.09.26, 14:00–15:30, Room: Runder Saal

Chair: Annette Kopp-Schneider

Authors: Monika Jelizarow [1]; Andrea Callegaro [2]; Björn Bornkamp [3]; Christian Stock [4]; Silvia Calderazzo [5]; James Travis [6]; Kit Roes [7, 8]; Nicky Best [9]; Andy Grieve [10]; Tim Friede [11]

Affiliations: [1] UCB Biosciences GmbH, Monheim am Rhein; [2] GSK Vaccines, Wavre; [3] Novartis Pharma AG, Basel; [4] Boehringer Ingelheim Pharma GmbH & Co. KG, Ingelheim, Germany, Ingelheim am Rhein; [5] German Cancer Research Center, Heidelberg; [6] Division of Biometrics II, Office of Biostatistics, Center for Drug Evaluation and Research, Baltimore; [7] Radboudumc, Nijmegen; [8] MA Methodology Working Party, Amsterdam; [9] GSK, UK; [10] King's College London, London; [11] Universitätsmedizin Göttingen, Göttingen

General outline and objectives

The FDA recently published a draft guidance titled "Use of Bayesian Methodology in Clinical Trials of Drug and Biological Products." This marks a major step toward broader acceptance of Bayesian methods in regulatory submissions, as the guidance is shifting towards a full acceptance of the Bayesian paradigm in the context of drug development and provides a roadmap for implementing Bayesian approaches in clinical trials. The draft guidance has attracted considerable attention and sparked discussions. Almost simultaneously, EMA issued a concept paper on the use of Bayesian methods in clinical development, indicating that attention to the topic is also shared by European regulators and is planned to develop into a future reflection paper. With the traditional paradigm potentially shifting, a discussion on the use of Bayesian methods for clinical trial design and analysis is timely.

Structure

14:00-14:50 Statements of panelists: James Travis, A Regulator's Perspectives on the Future of Borrowing and Bayesian Methods

Kit Roes, Level of evidence obtained from clinical trials that include external data from regulatory decision making perspective.

Nicky Best, Opportunities and Challenges of using Bayesian Methods to support Regulatory Decision Making

Andrew Grieve, Optimising the Trade-Off Between Type I and Type II Errors for Composite Hypotheses.

Tim Friede, Comprehensive cohort studies: Prospectively collecting and integrating randomized and non-randomized data

14:50-15:30 Panel discussion

Oral Sessions

Mon 28.09.2026

11:00–12:30

Frameworks for Clinical Research

Mon, 28.09.26, 11:00–12:30, Room: K2

Chair: André Scherag

ACRIBiS: Current Progress in Biosignal Data Integration and AI-Ready Infrastructure

11:00-11:15

Presenter: Lennart Graf

Authors: Lennart Graf [1]; Maximilian Oremek [2]; Sven Zenker [2]; Dagmar Krefting [1, 3]

Affiliations: [1] University Medical Center Göttingen, Dpt. of Medical Informatics, Germany, Göttingen; [2] University Hospital Bonn, Unit for Medical and Scientific Technology Development & Coordination (MWTek), Germany, Bonn; [3] DZHK (German Centre for Cardiovascular Research), Partner Site Göttingen, Germany, Göttingen

Introduction

The integration of electrocardiogram (ECG) data into clinical and research applications remains challenging, despite the central role that ECG data plays in cardiovascular diagnostics. Proprietary data formats and heterogeneous infrastructures limit standardized access and advanced analysis. The ACRIBiS project addresses these issues by creating a FAIR infrastructure that allows for the interoperable storage and processing of 12-lead ECGs in 15 university hospitals.

Methods

After requirements analysis to evaluate the characteristics of the site-specific infrastructure, we defined three ECG integration clusters: XNAT with DICOM storage, DICOMweb via OMI DICOM-RST, and Fleischhacker Mediconnect. We then implemented interoperable interfaces to AcuWave [1], a modular biosignal and time series processing platform, to enable harmonized data access and analysis. During the initial rollout phase, we successfully deployed the ECG integration software stack at the core sites. Next, we will integrate deep learning models to enable risk prediction. As a

proof of concept, a deep learning model was locally applied within the ACRIBiS software stack to estimate heart age from 21,799 ECGs from the PTB-XL database [2]. We calculated the mean and standard deviation of the predicted heart age and stratified them by sex.

Results

The ACRIBiS pipeline successfully processed all 21,799 ECGs from the PTB-XL database. The mean estimated heart age was 63.35 $\pm$ 15.88 years for ECGs from 11,354 male patients and 62.36 $\pm$ 17.96 years for ECGs from 10,445 female patients. In comparison, the mean chronological age was 60.20 $\pm$ 24.18 years for male patients and 65.56 $\pm$ 39.01 years for female patients. These results are consistent with findings from other studies [3].

Discussion

We successfully evaluated a state-of-the-art deep learning model within the ACRIBiS software stack in a local deployment. Due to its modular design, different deep learning models can be flexibly integrated and exchanged in the future. As the software is rolled out to additional sites, this software can be used for decentralized risk prediction based on biosignals and integrate them with the harmonized cardiovascular history within the ACRIBiS project.

References

[1] 10.3205/17gmds160

[2] 10.1038/s41467-021-25351-7

[3] 10.1038/s41746-024-01428-7

Optimization of Information Extraction via Layout Analysis in Medical Reports

11:15-11:30

Presenter: Lakisha Ortiz Rosario

Authors: Lakisha Ortiz Rosario; Stefan Lenz; Torsten Panholzer

Affiliations: Institute of Medical Biostatistics, Epidemiology and Informatics, University Medical Center of the Johannes Gutenberg University Mainz, Mainz

Introduction

Using large language models on medical reports is most efficient when only the sections relevant to a specific question are processed, as the processing time increases and answer quality decreases with the length of the input. Relevant sections can be identified via the document layout, which is designed for rapid human reading. While general-purpose layout parsers are versatile and can support these tasks, their broad scope can make them slow and computationally demanding.

Methods

We developed a software implementing a heuristic, density-based clustering algorithm for grouping text based on document layout, which works for both text-based and image-based PDFs. Results are compared against the layout parsing toolkit MinerU [1] regarding clustering quality (measured by V-measure [2] and over- and undersegmentation [3]), robustness against image distortions (measured by OCR accuracy), and runtime. The evaluation uses a dataset of 153 anonymised text-based PDFs of discharge letters. These were additionally transformed to image-based PDFs and augmented with synthetic scan artifacts. Gold standard text clusters were annotated manually for a subset of the dataset.

Results

The developed solution achieved a better balance between quality and efficiency compared to MinerU. While a comparable text extraction accuracy was achieved (maximum difference of 2%), the developed solution shows better results in clustering (V-Measure of 0.981 vs. 0.914) and a greater robustness against common scan artifacts like skewing. For image-based documents, a runtime advantage by a factor of 12 (avg. 0.05s vs. 50s per document) and for text-based documents by a factor 1000 was observed (avg. 5s vs. 62s per document).

Discussion

The results demonstrate that a specialized solution can yield high-quality results, enabling fast and local processing of medical documents. This allows to support efficient human-in-the-loop applications such as interactively using locally hosted large language models in clinical settings. The software is available under https://github.com/lortizro/PDFragmenter.

[1] Wang, Bin, et al. MinerU: An open-source solution for precise document content extraction. arXiv preprint arXiv:2409.18839, 2024. https://doi.org/10.48550/arXiv.2409.18839

[2] Rosenberg, Andrew Hirschberg, Julia. V-measure: A conditional entropy-based external cluster

evaluation measure. https://doi.org/10.7916/D80V8N84

[3] Antonacopoulos, Apostolos, et al. Icdar 2013 competition on historical newspaper layout analysis (hnla 2013). https://doi.org/10.1109/ICDAR.2013.293

A Lightweight Framework for Medical Named Entity Recognition and Relation Extraction

11:30-11:45

Presenter: Susanna Gordon

Authors: Sihan Wu [1]; Susanna Gordon [2]; Martin Boeker [2]; Luise Modersohn [3]

Affiliations: [1] Technische Universität München, München; [2] Technische Universität München (TUM), München; [3] Institute for Artificial Intelligence and Informatics in Medicine, Chair of Medical Informatics, Medical Center rechts der Isar, Technical University of Munich, Munich, Germany, München

Introduction

Clinical text in electronic health records contains valuable information for medical research. Named entity recognition (NER) and relation extraction (RE) are key tasks for structuring this information, but existing deep learning approaches are often computationally expensive.

Methods

We develop a lightweight, pipeline-based NER-RE system using neural networks with contextualized embeddings based on spaCy and Thinc. The impact of data preprocessing and hyperparameter tuning on NER is evaluated. RE is performed using a separate classifier on candidate entity pairs. Experiments are conducted on i2b2 and n2c2 datasets.

Results

The approach achieves strong NER performance after optimization. Better data quality or applying hyperparameter tuning improve performance comparably, but combining both yields only marginal gains additionally.

Discussion

Lightweight models are effective for NER but limited for RE due to error propagation. Our NER results also suggest that hyperparameter tuning can compensate for data quality issues.

The full paper of this abstract will be published in the GMDS series "Studies in Health Technology and Informatics" with doi: 10.3233/SHTI261002, URL:https://dx.doi.org/10.3233/SHTI261002

Medical Concept Normalization of German Clinical Expressions to SNOMED CT

11:45-12:00

Presenter: Helena Adam

Authors: Helena Adam [1]; Akhila Abdulnazar [1, 2]; Roland Roller [3]; Stefan Schulz [1]; Markus Kreuzthaler [1]

Affiliations: [1] Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Graz; [2] CBmed GmbH - Center for Biomarker Research in Medicine, Graz; [3] Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Berlin

Introduction. Clinical narratives in electronic health records frequently contain clinical expressions describing medical conditions. Their free-text format limits interoperability and automated processing. Medical concept normalization (MCN) addresses this challenge by mapping textual expressions to standardized terminology concepts, such as those from SNOMED CT. However, clinical language is characterized by abbreviations, spelling variants, and short forms, which complicate automatic normalization. Methods. In this work, we investigate the MCN of short German clinical expressions to SNOMED CT by comparing a direct large language model (LLM)-based normalization approach using GPT-5.4 with a hybrid retrieval approach that combines a medBERT.de bi-encoder for

embedding-based retrieval and retrieval-augmented generation (RAG) reranking using GPT-5 variants. Results. The LLM-only baseline achieves a Recall@1 of 0.235, Recall@3 of 0.297, and Recall@5 of 0.303. In contrast, the embedding-based bi-encoder retrieval approach achieves a Recall@1 of 0.681, a Recall@3 of 0.783, and a Recall@5 of 0.812. Incorporating RAG-based LLM reranking further improves Recall@1 to 0.771, while Recall@3 and Recall@5 reach 0.809 and 0.812, respectively. Discussion. The results demonstrate that combining domain-specific embedding-based retrieval with LLM-based reranking substantially improves MCN performance compared to a standalone LLM approach. This hybrid strategy supports robust semantic matching between German clinical expressions and standardized terminology concepts, facilitating structured representation and enabling more effective secondary use of clinical text.

The full paper of this abstract will be published in the GMDS series "Studies in Health Technology and Informatics" with doi: 10.3233/SHTI260995, URL:https://dx.doi.org/10.3233/SHTI260995

The development and validation of AI based systems to generate statistical analysis plans and analysis code

12:00-12:15

Presenter: Gordon Forbes

Authors: Hassan Jafari; Petrina Chu; Maximin Lange; Francesca Maher; Callum Glen; Ciel Burges; Oliver Pearson; Louise MacGregor; Samantha Cross; Meredith Martyn; Ben Carter; Richard Emsley; Gordon Forbes

Affiliations: King's College London, London

Introduction

AI is being touted as being touted as a revolutionary productivity tool which will automate many knowledge based tasks. Automating trial analysis would enable a larger number of trials and accelerate the the discovery of life-changing interventions. But AI adoption is not without risks, including hallucinations and the deskilling of the statistical workforce.

Methods

We developed and validated two different approaches to AI-assisted statistical analysis plan generation. 1) An automated prompting pipeline and 2) A human-in-the-loop retrieval augmented generation approach. The second approach can also produce statistical analysis code. Both approaches use commercially available large language models (LLMs).

The automated prompting pipeline generated entire statistical analysis plans by splitting the plan into sections, with a single prompt for each section. This was validated against 9 randomised trials, with AI-generated analysis plans scored for accuracy by trial statisticians against published content guidance.

The human-in-the-loop retrieval augmented generation approach focused on the outcomes and analysis methods sections of an analysis plan. Statisticians confirmed output at intermediate stages, and analysis descriptions and code were generated via reference to a curated database of methods. Validation was conducted in a similar way to the validation of the automated prompting pipeline.

Results

The automated prompting pipeline produced SAP drafts with high overall accuracy (77%-78%), with no difference in performance the three different LLMs ($p=0.79$), but with differences in performance by content type ($p < 0.001$). All models performed better on descriptive items (e.g., administrative details, trial design) than on items requiring statistical reasoning (e.g., modelling strategies, sensitivity analyses). Validation of the second approach is underway at the time of abstract submission.

Discussion

Our results suggest that AI systems can reduce the time taken to develop statistical analysis plans and code. Our automated prompting pipelines faced limitations in accuracy, especially in areas that require more statistical reasoning. The human-in-the-loop and retrieval augmented generation approach can improve accuracy whilst providing statisticians with involvement and control over key decisions. The tools presented can be explored at sapai.statsplan.com.

Trustworthiness of AI in Medicine

12:15-12:30

Presenter: Jacqueline Metsch

Authors: Jacqueline Metsch; Anne-Christin Hauschild

Affiliations: Institute for Predictive Deep Learning in Medicine and Healthcare, Gießen

Introduction

Artificial intelligence (AI), particularly deep neural networks, is increasingly applied in medicine due to its strong predictive performance. However, its integration into clinical practice is limited by a lack of transparency, as these models often function as "black-boxes," raising concerns about trust, ethics, and accountability. With regulatory frameworks such as the GDPR, HIPAA, and the EU AI Act, transparency and explainability of AI models is now obligatory. While explainable AI (XAI) aims to address these challenges, current methods often produce inconsistent results and lack standardized evaluation, making comparison and clinical adoption difficult. As the number of XAI methods grows, standardized benchmarking frameworks are needed.

Methods

To address this, we developed BenchXAI [1], a benchmarking framework that evaluates fifteen post-hoc XAI methods across clinical, imaging, signal, and biomolecular data. Evaluation robustness was improved through repeated validation over 100 randomized train-test splits (Monte Carlo Cross

Validation), reducing the influence of randomness and identifying stable relevance attributions. A sample-wise normalization approach enables consistent statistical analysis and clearer visualization. Additionally, BenchXAI incorporates diverse visualizations and statistical tests to support systematic comparison across models and data modalities.

Results

XAI performance was strongly task-dependent, but some methods showed consistent advantages. Integrated Gradients, DeepLift, DeepLiftShap, and GradientShap performed reliably across all data types. In contrast, methods such as Deconvolution, Guided Backpropagation, and LRP-α1-β0 showed inconsistent behavior on certain datasets. Furthermore, many XAI methods produced relevance attributions aligned with traditional interpretable models, such as logistic regression and random forests, indicating agreement between explainability techniques and state-of-the-art interpretable machine learning models.

Discussion

BenchXAI provides a systematic standardized framework for evaluating XAI methods across diverse biomedical applications. By combining repeated randomized testing with standardized evaluation, it enables the identification of robust methods while highlighting their limitations. Although no single method is universally optimal, several approaches demonstrate consistently strong performance. Overall, these findings emphasize the importance of rigorous benchmarking and careful method selection to support transparent, reliable, and trustworthy AI in medicine.

[1] Metsch, J. M., & Hauschild, A. C. (2025). Benchxai: Comprehensive benchmarking of post-hoc explainable ai methods on multi-modal biomedical data. Computers in Biology and Medicine, 191, 110124.

Longitudinal and Joint Models of Clinical Trajectories

Mon, 28.09.26, 11:00–12:30, Room: K3

Chair: Dimitris Rizopoulos

Hierarchical Clustering for Sparse Longitudinal Data: Extensions for Continuous and Ordinal Outcomes

11:00-11:15

Presenter: Shawn Han

Authors: Shawn Han [1]; Jessica Barrett [1]; Eleanor Winpenny [2]

Affiliations: [1] Medical Research Council Biostatistics Unit, University of Cambridge, Cambridge; [2] Mohn Centre for Children's Health and Wellbeing, School of Public Health, Imperial College London, London

Introduction

Longitudinal behavioural outcomes in population health are often collected using questionnaires at a limited number of waves, resulting in sparse and unbalanced data. Standard model-based longitudinal clustering approaches, such as latent class mixed models, face challenges in this setting, including unstable estimation with limited observations, violations of modelling assumptions and high computational burden. Hierarchical longitudinal clustering provides a useful alternative but introduces structural challenges, as early merging decisions are irreversible, making stability in initial stages critical. In addition, existing formulations primarily focus on continuous outcomes, whereas many longitudinal measures arising from questionnaires are ordinal. The aim of this work is to develop improved longitudinal clustering methods to address instability and allow use of ordinal and continuous longitudinal variables.

Methods

We have developed two methodological contributions targeting instability in sparse longitudinal clustering. For continuous outcomes, we introduce a roughness-penalized spline preprocessing strategy that stabilises merging among sparse trajectories using a P-spline-based roughness penalty on spline coefficients. For ordinal outcomes, we propose a new likelihood-based hierarchical clustering framework, where each cluster is represented by a proportional odds trajectory model and similarity is defined through likelihood loss under merging. To address non-identifiability in early stages, we introduce a surrogate likelihood formulation based on global threshold estimates, allowing clusters to contribute level information without imposing restrictive assumptions. A likelihood-based Calinski-Harabasz-type criterion is used to select the number of clusters and model complexity is adapted dynamically.

Results

Simulation studies based on the structure of Understanding Society Dataset, a large UK longitudinal household survey, show that the proposed methods improve stability of the clustering path and recovery of latent trajectory structures under extreme sparsity. The continuous extension improves robustness under both normal and gamma error distributions, while the ordinal framework yields stable and interpretable trajectory groupings. Both approaches are computationally efficient relative to latent class mixed models, achieving comparable or improved clustering performance without requiring restrictive distributional assumptions.

Discussion

These findings highlight early-stage instability as a key challenge in clustering sparse longitudinal data and demonstrate that appropriate regularization improves robustness and interpretability. The proposed framework provides a flexible and scalable approach for analysing questionnaire-based longitudinal data in life-course research.

Are you making the right choices in joint models for longitudinal and time-to-event data? A structured overview of functional forms.

11:15-11:30

Presenter: Felix Oppong

Authors: Felix Oppong [1, 2]; Dimitris Rizopoulos [1]; Thierry Gorlia [2]; Nicole Erler [3]

Affiliations: [1] Erasmus University Medical Center, Rotterdam; [2] EORTC HQ, Brussels; [3] UMC Utrecht, Utrecht

Introduction

Joint models for longitudinal and time-to-event data are widely used in clinical research, yet many of their most powerful features remain underutilised. In particular, the choice of functional form (association structure), which defines how the biomarker trajectory relates to the event risk, is often treated as a technical detail despite its central role in determining the model's assumptions and interpretation. As a result, researchers frequently rely on default specifications such as the current value association or shared random effects, without recognising that alternative formulations capturing other clinically relevant aspects of the biomarker exist and can lead to different interpretations and scientific conclusions.

Methods

We present an overview of functional forms that link longitudinal and time-to-event processes, highlighting established approaches, recent developments, and key conceptual differences between seemingly similar association structures. We discuss current value, slope, acceleration, cumulative exposure, delta (change), shared random effects, and variability-based formulations, and illustrate

their implementation using the JMbayes2 R package. We further discuss extensions that allow more flexibility in modelling, including interactions with covariates and time-varying effects.

Results

We show that different functional forms highlight different aspects of the biomarker process, leading to distinct interpretations even when applied to the same trajectory. Current value formulations capture the instantaneous biomarker level and reflect the subject's current biological state. Cumulative formulations instead summarise the biomarker history and quantify long-term exposure. Variability based formulations capture subject-specific fluctuations around the trajectory, identifying instability or irregularity in the biomarker as an additional prognostic signal. Together, these results emphasise that the choice of functional form determines which biological mechanism is being modelled and should be guided by the underlying scientific question.

Discussion

The choice of functional form has important practical implications for how joint model results are interpreted and communicated. Some formulations introduce practical challenges, for example when their association parameters do not share the same scale or units as the biomarker, making effect sizes difficult to interpret. Choosing the functional form that aligns with the scientific question is essential to ensure meaningful interpretation of the results of the model.

---

openstatsware - Cross-industry, Community-based Development of Open-source Statistical Software

11:30-11:45

Presenter: Wilmar Igl

Authors: Wilmar Igl [1, 2]; Alessandro Gasparini [3, 2]; Ya Wang [4, 5]; Daniel Sabanés Bové [6, 2]

Affiliations: [1] VERANEX, Uppsala, Sweden; [2] SIG openstatsware (EFSPI, PSI, ASA), Sheffield; [3] Red Door Analytics AB, Stockholm; [4] Gilead Sciences, Foster City, CA; [5] SIG openstatsware (EFSPI, PSI, ASA), Alexandria, VA; [6] RCONIS, Basel

Introduction

openstatsware was founded as a scientific working group of the American Statistical Association (ASA) in 2022 and became also a Special Interest Group (SIG) of the PSI and EFSPI organizations in 2024. Its main objectives are 1) to develop high-quality, open-source statistical software, 2) to promote good software engineering practices, and 3) to stimulate collaborations in the statistical community across organizations.

Methods

NA

Results

Software Software development of statistical methods is prioritized which are missing in the available open-source software, are of fundamental importance in biostatistics, and are relevant for pharmaceutical research. For example, active projects include the R/mmrm package, which sets a new benchmark for fitting Mixed Models for Repeated Measures (MMRM) in R. This project is complemented by the R/brms.mmrm package, which extends MMRM to Bayesian inference. The R/maicplus implements key statistical models for Health Technology Assessment (HTA) dossier submissions across multiple countries. Good Software Engineering Practices The workshop "Good Software Engineering Practice for R Packages" was held at 5 locations across 3 continents since 2023 and course materials were made publicly available. In 2024, the openstatsguide was published with recommendations for good software engineering practices. openstatsware also organized and chaired sessions dedicated to statistical software engineering at major statistics conferences including CEN2023 (Basel, CH) and PSI2025 (London). Collaborations openstatsware comprises 64 members from 42 organizations across North America, Europe and Asia (co-chairs: Daniel Sabanes Bove, RCONIS Ya Wang, Gilead Alessandro Gasparini, Red Door Analytics). The regional chapters for US/EU and EU/APAC meet monthly in video conferences to organize the working group, discuss ongoing projects, and share relevant news on statistical methodology and software engineering. Specific updates on software-related projects are publicly available in a central code repository at https://github.com/openstatsware. General information on the activities of openstatsware are published on the official website at https://openstatsware.org.

Discussion

Since August 2022, the cross-industry working group openstatsware has become an established community for the development of open-source statistical software for the pharmaceutical industry with substantial impact regarding its main goals of developing high-quality statistical software, promoting good software engineering practices, and connecting statisticians across organizations in academia, the industry and the public sector.

A Bayesian Joint Model for Linking Longitudinal Trajectories to Time-to-Event Functional Outcomes: Application to Duchenne Muscular Dystrophy

11:45-12:00

Presenter: Tasnim Hamza

Authors: Rossella Belleli [1]; Tasnim Hamza [2]; Keith R. Abrams [3]

Affiliations: [1] F. Hoffmann-La Roche Ltd, Basel; [2] F. Hoffmann-La Roche, Ltd, Basel; [3] Visible Analytics, Oxford

Introduction

Objective To develop and evaluate a Bayesian joint modelling framework that links individual-level longitudinal trajectories to time-to-event functional outcomes, and to quantify how treatment effects on the longitudinal process translate into effects on a clinically meaningful event. We illustrate the approach using data from the FOR-DMD trial (NCT01603407) which included 196 ambulatory boys with DMD aged 4-7 years. Methods We developed a Bayesian joint model linking repeated measurements of time to rise from floor (TTR, seconds) with the time to loss of ability to rise from floor (LoR). The longitudinal process was modelled using a linear mixed-effects model for log-transformed TTR, including treatment-specific time slopes, a quadratic time effect, and subject-level random intercepts and slopes. Baseline age and North Star Ambulatory Assessment (NSAA) score were included as covariates. The event process was specified as a proportional hazards model with a Weibull baseline hazard, modelling time from randomization to LoR. Dependence between the two processes was introduced via a current-value association structure, allowing the instantaneous hazard of LoR to depend on the latent TTR trajectory. Model estimation was performed in a fully Bayesian framework using rstanarm, jmbayes2 and JAGS. Sensitivity analyses explored alternative baseline hazard specifications, interaction terms, reduced covariate structures, spline-based hazards, and prior specification. Results The joint model identified a strong association between the underlying longitudinal trajectory and the risk of LoR (posterior mean association parameter 4.8 95% credible interval [CrI]: 3.7-6.1), indicating that worsening TTR trajectories substantially increase event risk. Treatment-related differences in longitudinal slopes translated into clinically meaningful differences in event hazards for example, an estimated 16% annual improvement in TTR (i.e. decrease) for daily versus intermittent corticosteroid regimens corresponded to a hazard ratio of 0.44 (95% CrI: 0.29-0.63) for LoR. Inference was stable across all sensitivity analyses.

Discussion

Conclusions The proposed joint modelling framework provides a coherent approach for linking longitudinal trajectories and time-to-event functional outcomes and for improving inference and prediction in clinical trials. The framework enables treatment effects on clinical milestones to be inferred indirectly through their impact on longitudinal trajectories, offering a flexible and robust approach applicable beyond the motivating DMD example.

Expected times in disease progression stages: a joint modeling approach based on continuous-time Item Response Theory

12:00-12:15

Presenter: Cécile Proust-Lima

Authors: Cécile Proust-Lima [1]; Viviane Philipps [1]; Federico Sirna [1, 2]; Alexandra Foubert-Samier [1, 2]

Affiliations: [1] Univ. Bordeaux, INSERM, Institut Bergonié, BPH, U1219, CIC 1401, Bordeaux; [2] Bordeaux Hospital, Bordeaux

Introduction

Expected timings under certain health conditions constitute key quantities in clinical practice to inform patients and caregivers on disease progression and help them anticipate future milestones. When milestones are defined by patient- or clinician-reported outcomes that are intermittently assessed with error at visits using Likert scales, multi-state models (Touraine et al, https://doi.org/10.1177/0962280 can't be used anymore. In Multiple System Atrophy (MSA), a rare neurodegenerative disease characterized by a progressive and fatal motor degradation, the Unified MSA Rating Scale assesses a variety of motor and non-motor symptoms, with five stages of disability (from completely independent to bedridden stage) as milestones. We develop an original approach based on the Item Response Theory (IRT) to compute sojourn timings in disease stages, which exploits the longitudinal item-response nature of health conditions.

Methods

The repeated items and terminal event (death) are jointly modeled in the shared-random-effect framework combining a continuous-time IRT model with a proportional hazard model (Saulnier et al, https://doi.org/10.1093/biomtc/ujaf163) the IRT submodel comprises a structural linear mixed model with item-specific cumulative probit models, and the association with the terminal event is defined through functions of the underlying health condition trajectory. Posterior estimands, namely expected durations in stages and life-expectancy until milestones, are derived from the model and its maximum likelihood estimates, with 95% confidence intervals (CI) obtained by parametric bootstrap. The methodology is implemented in the R package JLPM on the CRAN.

Results

The implementation has been validated in simulations and compared with the multi-state approach under different scenarios. Applied to the French MSA cohort (>700 patients, 15-year follow-up at maximum), we have shown a large variability in the expected time until reaching each disability stage depending on the prognostic factors e.g. the life expectancy until major dependency varies from 4.6 (95%CI=4.2,5.0) to 7.0 (95%CI=5.3,9.8) years after the first symptoms.

Discussion

The joint IRT model combined with posterior computations offers a useful tool to inform patients and caregivers. It integrates all the available information, exploits its underlying structure, including measurement error, and naturally handles interval censoring. In MSA, the expected durations have been summarized into an R-Shiny app co-constructed with clinicians.

Longitudinal Outcomes Ending in Death - The Relevance of While-Alive

12:15-12:30

Presenter: Dries Reynders

Authors: Dries Reynders; Els Goetghebeur

Affiliations: Ghent University - Department of Mathematics, Computer Science and Statistics, Ghent

Introduction

A primary task of statisticians is ensuring that questions asked are relevant and clearly defined. The estimand framework as adopted in ICH E9 R1 brought support on both fronts. Nevertheless, when targeting treatment effects on longitudinal outcomes, death is too often seen as a nuisance or even a source of incomplete data, rather than an inherent part of the whole.

Methods

A range of estimands has been developed to overcome the complexities that result from the non-existence of the longitudinal outcome after death. Hypothetical strategies, principal strata, separable effects and composite outcome approaches all target potentially relevant questions, but ignore at least one part of what is important to inform patients. From a patient-perspective we need an estimand that addresses an identifiable subpopulation, that acknowledges that participants may die and that death is distinct from being alive in a poor condition.

Results

Comparing outcomes among survivors at specific times t, does satisfy these requirements, but is often dismissed as 'not causal' because it conditions on being alive which is a post-randomization variable. Instead of seeing the outcome while alive as one dimension, we recommend to consider it as an inherently two-dimensional estimand with the survival contrast as the second component. Considering survival and the longitudinal outcome while alive is the most direct description of the actual impact of a treatment.

Discussion

We discuss two ways of estimation in an observational setting - double inverse probability weighting or outcome regression averaged over survivor specific survival chances - and their sufficient conditions. We compare assumptions with those required by alternative estimands and dive into the interpretation of results when treatment effects go in opposite directions on the two dimensions. This is illustrated by analysing simulated and case study data on quality of life.

Design and Analysis Choices in Clinical Trials

Mon, 28.09.26, 11:00–12:30, Room: K5

Chair: Elizabeth Williamson

A Minimal predictable randomization design for two- and multi-arm trials with equal or unequal allocations based on a unified assessment framework for treatment imbalance and selection bias risk

11:00-11:15

Presenter/Author: Wenle Zhao

Affiliations: Medical University of South Carolina, Charleston

Introduction

This presentation proposes a unified quantitative framework for evaluating treatment imbalance and selection bias risk in randomized controlled trials with two or multiple arm and equal or unequal target allocations. The framework integrates two complementary metrics: treatment imbalance, defined as the range of allocation-adjusted assignment residuals across treatment arms, and selection bias risk, quantified by the accuracy and frequency of allocation prediction under a generalized convergent prediction strategy adjusted for target allocation probability.

Methods

Based on this framework, the minimal predictable design, a randomization algorithm that minimizes selection bias risk subject to a pre-specified maximum tolerated treatment imbalance, is proposed. This new design is compared with established randomization methods applicable to multi-arm trials with unequal allocations, including the permuted block design, block urn design and mass-weighted urn design. Comparisons are conducted across multiple dimensions, including urn model operation, geometric allocation spaces, treatment imbalance boundaries, allocation predictabilities and selection bias risk.

Results

Results demonstrate that the minimal predictable design consistently achieves the lowest selection bias risk under equivalent imbalance constraints and is applicable across all trial settings.

Discussion

When evaluated on performance criteria, namely, controlling treatment imbalance and minimizing selection bias risk, the minimal predictable design consistently outperforms existing block-based and urn-based methods across a broad range of allocation settings. By introducing the smallest necessary deviation from complete randomization while enforcing a pre-specified imbalance boundary, MPD provides a practical and robust alternative for modern randomized controlled trials.

Stratified Randomization Incorporating Historical Control Data

11:15-11:30

Presenter: Masatoshi Shoji

Authors: Masatoshi Shoji [1]; Carolin Herrmann [2]; Isao Yokota [3]

Affiliations: [1] Department of Biostatistics, Graduate School of Medicine, Hokkaido University, Sapporo; [2] Mathematical Institute and Center for Digital Medicine, Faculty of Mathematics and Natural Sciences, Heinrich Heine University Düsseldorf, Düsseldorf; [3] Department of Biostatistics, Faculty of Medicine, Hokkaido University, Sapporo

Introduction

Randomized controlled trials (RCTs) are the gold standard for clinical trials, but are often infeasible due to limited sample size, high costs, or lengthy durations, particularly in rare diseases. Hybrid controlled trial designs are one option to address these challenges by combining an internal control arm with historical data. However, existing methods, e.g. propensity score matching (PSM), adjust for confounding only at the analysis stage. This, however, limits power gains and risks covariate imbalance when covariate distributions differ between historical and internal data.

Methods

We propose a stratified block randomization method for hybrid control trials. Strata are defined by risk factors, and a 1:1 treatment-to-control ratio is ultimately achieved within each stratum by combining internal and historical controls. To this end, the allocation ratio for each stratum is determined based on the number of subjects in the historical data. Internal participants are allocated to the treatment arm at a higher ratio until the available historical data in each stratum are fully used thereafter, allocation proceeds at a 1:1 ratio. More precisely, we consider two approaches: a random sampling from historical participants, and a weighting method that reflects the number of historical participants. We compared these two proposed methods with PSM in terms of power, type I error, and sample size via Monte Carlo simulations under various heterogeneity scenarios between historical and internal data.

Results

Simulations demonstrated that the newly proposed methods maintained type I error rates unless there was heterogeneity in the outcome model. Compared with PSM, the newly proposed methods achieved higher power and smaller sample size when covariate distributions were imbalanced between historical and internal data, and this advantage became more pronounced as the degree of imbalance increased. In particular, the weighting method generally showed the highest power and lowest type I error among the compared methods.

Discussion

The proposed method is novel in that it incorporates historical data at the allocation stage, rather than only at the analysis stage as in PSM. This enables more efficient use of historical data and improves power while maintaining type I error control.

Sample Size Determination for Superiority Trials with Survival Outcomes under varying patterns of proportional hazards: Choice of Summary Measure and Truncation Time

11:30-11:45

Presenter: Tianyuan Gu

Authors: Tianyuan Gu [1]; Yu Yang Soon [2]; Joseph Wee [3, 4]; Bee Choo Tai [1, 5]

Affiliations: [1] Saw Swee Hock School of Public Health, National University of Singapore and National University Health System, Singapore, Singapore; [2] Department of Radiation Oncology, National University Cancer Institute, National University Hospital, Singapore, Singapore; [3] Division of Radiation Oncology, National Cancer Centre Singapore, Singapore, Singapore; [4] Duke-NUS, Graduate Medical School, Singapore, Singapore; [5] Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

Introduction

In superiority trials with time-to-event outcomes, the hazard ratio (HR) from the Cox proportional hazard (PH) model remains the standard summary measure, but its interpretation and efficiency may be compromised when the PH assumption does not hold. Alternative HR-based, difference in survival (DS)-based, and mean survival time (MST)-based measures have been proposed, yet their relative performance across truncation times and in sample size determination remains unclear.

Methods

Intensive simulation was conducted assuming survival times following Piecewise Exponential distribution under a balanced design with sample sizes ranging from 100 to 800. PH (HR = 0.5), diverging (HR = 1 and 0.3 thereafter), crossing (HR = 1.5 and 0.5 thereafter), and diminishing treatment effects (HR = 0.4 initially, increasing to >1 over time) were assumed. HR-, DS- and MST-based measures were compared across truncation times and censoring levels (high and low). HR-based measures included the Cox, Yang-Prentice (YP), average HR models, and unweighted and weighted log-rank (LR) tests. DS-based measures were obtained from the Cox model and Kaplan-Meier estimator. MST-based measures included the restricted MST, window MST, weighted Kaplan-Meier, and integrated average difference. Statistical power was evaluated across truncation times and windows using a two-sided α of 0.05, and the minimum sample size required to achieve 80% power was also determined. The performance of these methods was further illustrated based on four randomised clinical trials.

Results

No single summary measure performed best across scenarios. Cox (HR) and YP detected earlier effects at shorter truncation times under settings of PH, diverging and crossing. LR-related methods performed well when their weighting structure aligned with the period of largest treatment effect. MST-based measures were more robust at longer truncation times and diminishing treatment effects. DS-based measures performed best when truncation aligned with maximal survival separation. Power was higher for truncation windows that aligned with the period of large treatment effect.

Discussion

The choice of summary measure and truncation parameter materially affects both power and sample size under non-PH. At the design and analysis stages, selection of summary measure should be guided by the anticipated pattern of treatment effect, clinically relevant follow-up horizon and interpretability of the estimand.

WREstimates: An R package for win ratio sample size and power calculations

11:45-12:00

Presenter/Author: Autumn Johnson

Affiliations: University of Galway, Galway

Introduction

The win ratio has great potential for determining the overall efficacy of treatments and therapies in clinical trials, as its ability to hierarchically account for multiple endpoints provides a holistic metric of the treatment effect. For the win ratio to become a prominent and reliable statistical method beyond cardiovascular disease, a straightforward approach to study design, particularly for power and sample size determination, is needed.

Methods

Using an appropriate method for determining these metrics is vital to ensure the validity of the results obtained in a study.

Results

The WRestimates package provides easy-to-use functions for determining the required sample size and power for studies implementing the win ratio.

Discussion

These methods enable the calculation of sample size and power based on estimands or pilot data, thereby eliminating the need for complex simulation-based methods that require numerous assumptions about the data.

The full paper of this abstract will be published in the GMDS series "Studies in Health Technology and Informatics" with doi: 10.3233/SHTI261015, URL:https://dx.doi.org/10.3233/SHTI261015

Practical considerations for the analysis of cluster randomised trials with interval censoring

12:00-12:15

Presenter: Clemence Leyrat

Authors: Clemence Leyrat [1]; Agnes Caille [2, 3]; John Bradley [1]

Affiliations: [1] London School of Hygiene & Tropical Medicine, London; [2] Université de Tours, Université de Nantes, INSERM, SPHERE U1246, Tours; [3] INSERM CIC 1415, CHRU de Tours, Tours

Introduction

Although cluster randomised trials (CRTs) reporting a time-to-event outcome are common, the exact time at which the outcome of interest happens is not always known, but instead only a time window during which the outcome occurs is available. This is interval censoring is particularly common in CRTs that randomise a few large geographical areas with outcomes recorded during planned surveys. Interval censoring can be handled with complementary-loglog model, and for CRTs, there are published examples of its use with the addition of cluster-specific random effects. However, to date, there is (i) no evaluation of this approach for typical CRTs with few clusters, a situation in which mixed-models produce inflated type I error rates, (ii) no comparison with alternative methods, or (iii) no clarification of the estimand targeted.

Methods

We conducted Monte-Carlo simulations to compare the performance of a cluster-level analysis of survival probabilities, a Cox model with shared frailty, a c-loglog mixed-model, and GEE with a c-loglog distribution. We varied the number of intervals, the strength of clustering, the number and size of clusters, and the censoring rate. For the individual-level methods, we investigated the performance of small sample corrections (between-within method for mixed-models, Fay-Graubard for GEE) for type I error control. Effect measures were expressed as hazard ratios (HR) to mirror current practice, but differences in survival probabilities were also estimated.

Results

All the methods led to unbiased HR estimates, but greater precision was obtained using a cluster-level approach or a frailty model. Although convergence issues were <0.5% for most methods, a singular fit of the c-loglog mixed-model occurred in half of the simulations when clustering was weak (CV=0.1). A standard implementation of individual-level methods led to inflated type I error rates, as high as 13%, but appropriate corrections (between-within and Fay-Graubard corrections) performed well across simulations.

Discussion

We showed that the commonly used method to handle interval censoring in CRTs can lead to inflated type I error rates, but easy-to-implement small sample corrections perform well. We also clarified which specific estimand is targeted by each method, so that applied researchers can choose the method that aligns with their research question.

Is control of type I error rate needed in Bayesian clinical trial designs?

12:15-12:30

Presenter/Author: Elja Arjas

Affiliations: University of Helsinki, Helsinki

Introduction

Practical employment of Bayesian trial designs is still rare. Even if accepted in principle, the regulators have commonly required that such designs be calibrated according to an upper bound for the frequentist type I error rate. This represents a conceptually inconsistent hybrid methodology, where important advantages from following the Bayesian principles are lost. A practical consequence is that, in sequential designs, all interim looks need to be accounted for in advance and then have an inflating effect on type I error rate.

Methods

To present an alternative approach, we consider the prototype case of a 2-arm superiority trial with dichotomous outcomes. The design is adaptive, using error control based on sequentially updated posterior probabilities, to conclude efficacy of the experimental treatment or futility of the trial. As gatekeepers for a proposed design, the regulators have the main responsibility in determining acceptable levels for the parameters of the control of false positives, whereas the trial sponsors and investigators will have a natural role in specifying the criteria for stopping the trial due to futility. It is suggested that the traditional frequentist operating characteristics in the design, type I and type II error rates, be replaced, respectively, by Bayesian criteria called False Discovery Probability (FDP) and False Futility Probability (FFP) both terms correspond directly to their probability interpretations. Importantly, sequential error control during the data analysis based on posterior probabilities will satisfy these numerical criteria automatically, without need of preliminary computations before the trial is started. The method contains the option of applying a decision rule for terminating the trial early if the predicted costs from continuing would exceed the corresponding gains.

Results

Some computer simulation results are shown for illustration.

Discussion

The presentation ends with a general methodological discussion.

Reference: E. Arjas, D. Gasbarra (2026). Is control of type I error rate needed in Bayesian clinical trial designs? arXiv:2312.15222v5 [stat.ME]

Missing Data and Measurement Error

Mon, 28.09.26, 11:00–12:30, Room: N1

Chair: Jonathan Bartlett

Analysis of randomised controlled trials with missing data using the 'multiple imputation doctor' (midoc) software

11:00-11:15

Presenter: Elinor Curnow

Authors: Elinor Curnow [1]; Rosie Cornish [1]; Jon Heron [1]; Kate Tilling [1]; James Carpenter [2, 3]

Affiliations: [1] University of Bristol, Bristol; [2] London School of Hygiene & Tropical Medicine, London; [3] University College London, London

Introduction

Randomised controlled trials (RCTs) often have missing data. Choosing the most appropriate missing data strategy can be complex, involving multiple decisions. Although there are some guidelines for best practice, these are often hard to follow or outdated, e.g. developed prior to the ICH E9(R1) addendum. The multiple imputation doctor ('midoc') software is a guidance system for analysis with missing data. It incorporates expert, up-to-date methodology to help researchers choose the most appropriate analysis approach when some data are missing. We have extended our software to address particular issues that occur when there are missing data in RCTs. The midoc software can help the user identify which missing data method is most appropriate for their trial, where this will depend on the estimand of interest, intercurrent event (ICE) strategy, available information, and reasons for missing data.

Methods

We show how the midoc software can be used to both plan and implement analyses of RCTs with missing data. We demonstrate both our R package and web-based (code-free) application, using examples from RCTs of i) treatment for Parkinson's disease and ii) antidepressants.

Results

We illustrate how midoc can help with common problems in analyses of RCT data, including:

How to choose a missing data approach that is aligned with the estimand of interest How to depict assumptions about the causes of missing data - we show how directed acyclic graphs (DAGs) can be used to guide the missing data approach When complete case analysis, multiple imputation, or repeated measures analysis may be the preferred missing data approach How to conduct sensitivity analyses to explore the impact of different assumptions about the missing data

Discussion

The midoc software guides researchers through their analyses of RCTs with missing data, advising on the most appropriate missing data approach and how best to perform it, based on the user's trial design and missing data structure. By providing documented decisions and code, use of midoc will increase reproducibility and transparency of RCT analyses.

To Impute or Not to Impute? The Role of Bilateral Correlation on Handling Missing Data

11:15-11:30

Presenter: Pau Esteve Bricullé

Authors: Pau Esteve Bricullé [1, 2]; Nuria Perez Alvarez [1]; Jimena Fernández Carneado [3]; Guadalupe Gómez Melis [1]

Affiliations: [1] Universitat Politècnica de Catalunya, Barcelona; [2] BCN Peptides S.A., Barcelona; [3] BCN Peptides S.A., Sant Quintí de Mediona

Introduction

Bilateral clustered data, common in ophthalmic and other paired-organ clinical trials, present unique analytical challenges in the presence of missing data. Strong inter-unit correlation may render complex imputation algorithms unnecessary or even harmful. While likelihood-based models and multiple imputation are widely recommended over naive approaches, their relative performance as a function of paired correlation strength remains poorly characterized. This study evaluates whether the decision to impute should be guided by the nature of the bilateral variable itself.

Methods

We conducted a non-parametric simulation study based on complete-case data from a Phase II/III trial assessing diabetic ocular impairment over two years. Missingness was manually introduced under MCAR, MAR, and MNAR mechanisms, incorporating varying severities and complex structural patterns, including multi-mechanism scenarios. We evaluated the performance of several approaches: naive estimands, nested mixed models without explicit imputation, multiple imputation by fully conditional specification (under both congenial and uncongenial clustering assumptions), and expectation-maximization with bootstrapping. Across 200 replications, we assessed performance based on bias, RMSE, and 95% CI coverage/width, distinguishing between a moderately correlated functional variable and a highly correlated anatomical variable.

Results

The likelihood-based nested model consistently maintained near zero bias and nominal coverage regardless of missingness mechanism or severity for both outcomes. Under moderate correlation, imputation congeniality was relevant subject-clustered MICE increasingly outperformed its uncongenial counterpart as missingness grew. The EMB algorithm maintained optimal coverage across

all moderate correlation scenarios. For highly correlated variables, MICE performed adequately at lower missingness but deteriorated under compounded mechanisms, with uncongenial MICE coverage falling below 50% and EMB below 80%. Subject-level averaging proved the most robust under high correlation and severe missingness, maintaining 95% coverage even under the most adverse conditions.

Discussion

The decision to impute must be guided by bilateral correlation strength and missingness patterns on each variable. For highly correlated variables, likelihood approaches without explicit imputation are the safest default standard multiple imputation may paradoxically worsen inference. For moderately correlated variables, congenial imputation preserves validity, particularly when covariate missingness is present. These findings aim to yield practical, correlation stratified recommendations for handling missing data in paired clinical trials.

Addressing Missing Data in Menstrual Health Research: A Pattern-Based Imputation Framework from the MeMaF Study

11:30-11:45

Presenter: Thomas Gärtner

Authors: Thomas Gärtner [1]; Hannah Neumayer [1]; Rebecca Paprott [1]; Daniel Pach [2]; Stefan Konigorski [1]; Lorena Hafermann [3]

Affiliations: [1] Hasso Plattner Institut, Potsdam; [2] Institut für Sozialmedizin, Epidemiologie und Gesundheitsökonomie, Charité Universitätsmedizin Berlin, Berlin; [3] Institute of Biometry and Clinical Epidemiologie, Charité Universitätsmedizin Berlin, Berlin

Introduction

In medical research, self-tracked health data, such as daily symptom diaries or device-based measurements, are increasingly collected but often affected by complex missingness patterns. In particular, irregular and symptom-driven reporting can lead to non-standard missingness patterns challenging missing-at-random (MAR) assumptions. We consider the MeMaF study, a nested randomized controlled trial evaluating a new hybrid multimodal care model for menstrual pain in women aged 16-24 years without an endometriosis diagnosis. MeMaF uses daily digital diaries to assess menstrual pain days within 28-day time windows. In this setting, missingness is closely linked to symptom occurrence, complicating standard approaches to handling incomplete data.

Methods

In the MeMaF study, a pattern-based approach was used to impute missing entries for menstrual pain, including the following patterns within 28-day time windows: (1) In windows that had 4-10 nonmissing entries with entries around days of menstrual bleeding, suggesting symptom-triggered reporting consistent with missing-not-at-random (MNAR), no pain was imputed on all other days in

the window. (2) In windows with more arbitrary missingness patterns but sufficient observations, pain days were imputed based on the observed proportion of symptomatic days. (3) In windows with <4 diary entries, missing values were imputed using multiple imputation by chained equations under a MAR assumption on an aggregated monthly level. We validated the proposed pattern-based approach with respect to the bias in estimating treatment effects, using a simulated dataset that encompassed three scenarios of missing data: (1) missingness conditional on menstrual bleeding (MNAR) (2) missingness based on the duration of the study (MAR) and (3) MCAR.

Results

The simulation results indicate that no single imputation method is universally optimal, motivating a carefully designed imputation approach. In the MeMaF study, adherence to daily diary reporting varied among participants and across time points, resulting in heterogeneous patterns of missingness. The predefined strategy outlined above enabled systematic classification of diary completeness and derivation of all primary outcomes at the monthly level.

Discussion

Combining pattern-based classification with deterministic and model-based imputation can provide a framework for handling missing data in diary studies with symptom-triggered reporting.

Missing data strategy in bioavailability studies at Danone Research & Innovation

11:45-12:00

Presenter: Luca Carlini

Authors: Luca Carlini [1]; Floor Van Oudenhoven [1]; Sophie Swinkels [1]; Nicole Erler [2]

Affiliations: [1] Danone Research & Innovation, Utrecht; [2] UMC Utrecht, Utrecht

Introduction

Missing samples are frequent in bioavailability studies with dense time courses. In our current practice at Danone Research & Innovation, if a value is missing near the expected peak of the concentration curve, the whole visit is deemed unevaluable and a replacement participant is recruited otherwise, simple interpolation may be used. This can reduce power, introduce bias for Cmax (maximum/peak concentration), Tmax (time to peak), and iAUC (incremental area under the concentration-time curve), increase cost and timelines, and add burden for volunteers and test centers. Our aim is to keep as many visits analysable as possible without weakening inference.

Methods

We are developing a Bayesian imputation workflow that models the concentration-time profile jointly across time within visit and product, implemented in R (e.g., JointAI). The model borrows information from all remaining observations, product/visit indicators, and design covariates. Added uncertainty due to the missing values is automatically included in the posterior distribution. Building on internal

SAP practice for bioavailability endpoints, we embed clear decision rules describing when imputation can support analysis versus when participant replacement remains necessary.

Results

We will present simulations calibrated to internal trial designs (sampling grids, variability, effect sizes) quantifying bias, precision, coverage, and Type I error for Cmax, Tmax, and iAUC. Scenarios vary missingness mechanisms, missing at random (MAR) and missing not at random (MNAR), timing (near-peak vs off-peak), and extent (subjects/time points). We also report practical diagnostics, operating characteristics of the decision rules, and guidance on how many, and which, missing samples can be tolerated while avoiding replacements.

Discussion

A pre-specifiable Bayesian imputation framework can protect inference on dense-sampling bioavailability endpoints while reducing avoidable waste, cost, and site/participant burden, with transparent rules for when imputation suffices and when replacement is warranted.

Simulating the effect of imputing missing pharmacokinetic sampling time points on bioequivalence assessment in typical early phase clinical trials

12:00-12:15

Presenter: Jakob Winkler

Authors: Jakob Winkler; Markus Waser

Affiliations: Staburo GmbH, Munich

Introduction

Pharmacokinetic (PK) analysis is essential in clinical trials to characterize a drug's profile, which provides the empirical basis for establishing safe and effective dosing regimens and for assessing bioequivalence between two drugs. Regulatory guidelines provide strict bioequivalence ranges based on the comparison of the geometric means of PK parameters such as the area under the time-concentration curve (AUC). In practice, missing actual sampling time points are often replaced by planned time points. This approach, however, introduces a potential bias which might compromise the accurate calculation of the AUC and the bioequivalence conclusions based on it. The objective of this research is to quantify this effect based on varying simulated scenarios for a typical bioequivalence trial.

Methods

We simulated 500 bioequivalence trials for two oral drugs in a typical 2x2 crossover design with N=12 trial participants. PK profiles were generated using a one-compartment model with varying half-lifes and for geometric mean AUC ratios within the standard equivalence range 80%-125%. Observed

sampling times within each trial were perturbed around the planned schedule using uniformly distributed deviations within a +/-10 min window. Missingness was then introduced at the observation level under a missing-not-at-random mechanism, with higher probabilities for time points close to the time of maximum concentration tmax, following a bell-shaped distribution. This was repeated 100 times. Missing sampling times were imputed by the planned sampling times and the bioequivalence results compared to those using the actual sampling time points.

Results

For a true geometric mean ratio close to the lower equivalence bound (80%), a short half-life of approximately 1.2 h, and a timing deviation of maximum +/-10 min, missingness resulted in a false non-equivalence rate of 2-6% across trials. The variation of the false non-equivalence rate amounted to approximately 2-3%. Under a missing-at-random mechanism with the same proportion of missing observations, about 1% fewer false rejections were observed.

Discussion

Missing time points potentially influence the results of bioequivalence trials to a relevant extent. To account for this, simulating different scenarios of missingness as sensitivity analysis might be considered to strengthen the trial claims - especially when the trial results are close to the bioequivalence bounds.

Assessing the Impact of Measurement Error on Fractional Polynomials and Spline Estimates of Non-Linear Dose-Response Functions: A Simulation Study

12:15-12:30

Presenter: Steve Ferreira Guerra

Authors: Steve Ferreira Guerra [1]; Laurence Freedman [2, 3]; Paul Gustafson [4]; Willi Sauerbrei [5]; Frank E. Harrel Jr. [6]; Michal Abrahamowicz [7]

Affiliations: [1] Centre de Recherche du Centre Hospitalier de l'Université de Montréal (CRCHUM), Montréal; [2] Data and Analytics Branch, Sheba Medical Center, Tel Hashomer; [3] Information Management Services Inc., Rockville; [4] Department of Statistics, University of British Columbia, Vancouver; [5] Institut für Medizinische Biometrie und Informatik, Universitätsklinikum Freiburg, Freiburg; [6] Department of Biostatistics, Vanderbilt University School of Medicine, Nashville; [7] Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal

Introduction

Accurately characterizing potentially complex, non-linear relationships between continuous risk or prognostic factors and outcomes is central to modern epidemiological and biostatistical analyses. Yet, in practice, these predictors are frequently measured with error, and the impact of such measurement error (ME) on flexible non-linear estimation remains largely unexplored.

Methods

We conducted an extensive simulation study to evaluate the impact of classical random ME on non-linear exposure-outcome relationships estimated using logistic regression. Data were generated under a range of realistic scenarios varying sample size, predictor distributions, and underlying functional forms. ME was introduced according to a classical additive error model with increasing variance. For each simulated dataset, we estimated the dose-response function using fractional polynomials and three spline-based approaches (B-splines, natural splines, and penalized splines). Performance was assessed using graphical comparisons of estimated versus true functions and quantitative metrics capturing both bias and variability.

Results

We found that ME generally results in an attenuation of the true strength of the relationship, with increasing ME leading to artificially flatter and more linear estimated functions. However, random ME could also induce local distortions, including spurious curvature, particularly in the tails of the predictor distribution. Interestingly, increasing ME may also sometimes reduce the local variance of the estimates, especially in the tails where it may artificially augment the number of observations. In some settings, increasing ME reduced the variability of estimates, especially in sparse regions, by effectively increasing local data density. The magnitude and nature of these effects depended strongly on the varying parameters, especially underlying predictor distribution.

Discussion

Overall, our findings highlight the complex and sometimes counterintuitive impact of ME on flexible estimation of exposure-outcome relationships. Beyond global attenuation, ME may alter both the shape and variability of estimated functions in ways that depend on data characteristics. These results underscore the need for greater awareness of ME in non-linear modeling and for the development, validation, and application of methods to mitigate its impact.

Digital Health Tools and Engagement

Mon, 28.09.26, 11:00–12:30, Room: N2

Chair: Simona Dobre

From theory to practice: Operational stakeholder management frameworks in healthcare - A Scoping Review

11:00-11:15

Presenter: Henrike Räther

Authors: Henrike Räther [1]; Ekaterina Kutafina [2]; Barbara Eichhorst [1]; Oya Beyan [3]

Affiliations: [1] Department I of Internal Medicine, University of Cologne, Germany, Cologne; [2] Institute for Biomedical Informatics, Faculty of Medicine and University Hospital Cologne, University of Cologne, Germany, Cologne; [3] Institute for Biomedical Informatics, Faculty of Medicine and University Hospital Cologne, University of Cologne, Germany, Köln

Introduction

Stakeholder management (ShM) is essential for healthcare projects, particularly for medical data platform development, which involves diverse stakeholders and complex socio-technical challenges spanning ethical, legal, technical, and human and organizational dimensions. While theoretical frameworks exist, their practical implementation often remains unclear. Operational ShM frameworks that provide actionable guidance are critical to ensure that project managers can effectively engage stakeholders to improve project outcomes.

Methods

A scoping literature search in Google Scholar and Scopus screened the top 30 results from each database for studies presenting operational ShM, stakeholder analysis (ShA), or engagement (ShE) frameworks in healthcare research or institutions. Ten studies were included.

Results

Frameworks varied in scope, methods, and completeness four focused on ShA, four on ShE, and two covered the full ShM process. Only one study provided a comprehensive guideline spanning all steps and including continuous stakeholder management.

Discussion

Operational ShM frameworks in healthcare remain scarce, highlighting the need to develop a practical, actionable framework providing clear guidance for stakeholder management. Such a framework is critical for effectively coordinating diverse stakeholders and supporting project managers in implementing robust, context-sensitive digital health infrastructures.

The full paper of this abstract will be published in the GMDS series "Studies in Health Technology and Informatics" with doi: 10.3233/SHTI260990, URL:https://dx.doi.org/10.3233/SHTI260990

Technical Barriers and Facilitators to Adherence and Patient Engagement in Digital Patient-Reported Outcome Collection: An Umbrella Review

11:15-11:30

Presenter: Lilly Sophia Brandstetter

Authors: Lilly Sophia Brandstetter [1, 2]; Miriam Schlüter [1, 2]; Steffi Jírů-Hillmann [2]; Carsten Vogel [3, 2]; Alizé Rogge [4]; Rebecca Mukowski-Kickhöfel [5, 4]; Lisa Rebecca Otto [5, 4]; Annika Braun [6]; Paul Hüsing [6]; Philipp Breitfeld [7]; Matthias Rose [5, 4]; Peter U. Heuschmann [8, 2]; Rüdiger Pryss [8, 2]

Affiliations: [1] University Hospital Würzburg, Institute for Medical Data Sciences, Würzburg; [2] Julius-Maximilian University Würzburg, Institute for Clinical Epidemiology and Biometry, Würzburg; [3] University Hospital Würzburg, Institute for Medical Data Sciences, Würz; [4] Department of Psychosomatic Medicine, Charité - Universitätsmedizin Berlin, Berlin; [5] Center for Patient-Centered Outcomes Research (CPCOR), Charité - Universitätsmedizin, Berlin, Berlin; [6] Department of Psychosomatic Medicine and Psychotherapy, University Medical Center Hamburg-Eppendorf, Hamburg; [7] Center for Experimental Medicine, Institute for Applied Medical Informatics University Medical Center Hamburg-Eppendorf, Hamburg; [8] Institute of Clinical Epidemiology and Biometry, Julius-Maximilians University Würzburg, Würzburg

Introduction

Patient-reported outcome measures (PROMs) are becoming increasingly important for the evaluation of treatment outcomes, promoting a patient-centered approach to clinical decision-making. However, the completion or response rates of PROMs by patients in clinical practice decrease over time. Multiple interventions or instruments have been employed to promote patient engagement with PROMs. The aim of this systematic umbrella review was to identify and summarise existing reviews evaluating technical barriers and facilitators of patient adherence to and engagement with PROMs.

Methods

The systematic umbrella review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. The electronic databases PubMed and Web of Science were searched on 2 October 2025. All types of review articles reporting on barriers and facilitators of patient adherence and engagement to PROMs published to this point were eligible. A narrative synthesis of the reviews' uncombined results was performed. Risk of bias assessment for individual review articles was performed using the AMSTAR 2 tool. The study protocol was registered at the PROSPERO database (CRD420251160007).

Results

A total of 11 review articles met the inclusion criteria and were included in the analysis. 41 barriers and facilitators of patient adherence and engagement to PROMs could be identified and were grouped into technological factors (e.g. accessibility, usability, reminders, interactivity), content-related factors (e.g. perceived usefulness, appropriateness of PROMs), monitoring-related factors (e.g. patient-provider communication, alerts/ feedback), and training (e.g. patient education/ skills).

Discussion

Different types of barriers and facilitators of patient adherence and engagement to PROMs were identified. Improving accessibility and usability of PROMs and implementing methods improving interactivity and patient-provider communication may be important strategies to improve patient adherence and engagement to PROMs. However, future studies should focus on evaluating the effect of technical methods beyond reminder systems, such as interactivity and gamification in ePRO tools.

MeLTSy: A MeSH-based system for linking medical terminology with laypeople synonyms

11:30-11:45

Presenter: Caroline Sautier

Authors: Caroline Sautier; Martin Wiesner

Affiliations: Heilbronn University, Heilbronn

Introduction

Research indicates that approximately 44% of the German population has difficulty understanding health information correctly [1]. Although there are a few online dictionaries that explain medical terminology for terms from specific fields, the majority of these resources require basic medical knowledge. Despite this offering, there is still no uniform, professionally reviewed thesaurus that systematically links medical terms with colloquial language.

For these reasons, it seems sensible to develop a supplementary directory of easily understandable terms and an associated information system to extend the existing MeSH specialist thesaurus.

Methods

Since 2020, ZB MED is responsible for translating the Medical Subject Headings (MeSH) into German, which is accessible and available free of charge in multiple formats [2].

To map between technical and layperson language, specifically structured and verified online sources were used: Wiktionary which maps medical terms with synonyms in technical and everyday language and their hierarchical relationships. Similarly, Wikidata and Duden supported the selection and validation of terms. In a multi-stage validation process, all three data sources were consulted to find a consensus on the most suitable synonym.

Results

The MeSH-based layperson terminology system (MeLTSy), which automatically provides verified synonyms for medical terms, was implemented in Java. A PostgreSQL database was used to record the relationships between technical and lay language, enabling consistent and traceable assignment. Additionally, a programming interface (API) was implemented that allows specific queries for synonyms and their German or English technical terms. Similarly, technical terms can be checked for potential synonyms.

Discussion

MeLTSy provides a technical foundation for: (a) the editing and proofreading of medical texts, (b) the simplification and translation for laypeople, and (c) the ad hoc lookup of synonyms. This way, the presented system combines medical terminology with colloquial German language to facilitate access to health information for most readers.

References

[1] Kröger J, Maaß C. Mangelnde Verständlichkeit durch Fachsprache in Gesundheitsinformationen zu chronischen Erkrankungen - eine qualitative Korpusanalyse. Prävent Gesundheitsförderung 202419:497-503. https://doi.org/10.1007/s11553-024-01124-0.

[2] Langnickel L, Baum R, Wollnik-Korn G, et al. The future of German MeSH: a new semi-automatic translation process and new services for search and annotation. GMS Med Inform Biom Epidemiol 202016:Doc07. https://doi.org/10.3205/mibe000210.

AI-Supported Chat System to Assist Telemedicine Staff in Monitoring Kidney Transplant Patients

11:45-12:00

Presenter: Roland Roller

Authors: Roland Roller [1]; Ajay Madhavan Ravichandran [2]; Oliver Amft [3]; Aditya Kumar [3]; Mario A. Cypko [3]; Bilgin Osmanodja [4]; Klemens Budde [5]

Affiliations: [1] Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Berlin; [2] German Research Center for Artificial Intelligence (DFKI), Berlin, Germany, Berlin; [3] Hahn-Schickard, Freiburg; [4] Department of Nephrology and Medical Intensive Care, Charité - Universitätsmedizin Berlin, Corporate Member of Free University of Berlin, Berlin Institute of Health, Humboldt-University of Berlin, Ber, Berlin; [5] Department of Nephrology and Medical Intensive Care, Charité - Universitätsmedizin Berlin, Corporate Member of Free University of Berlin, Berlin Institute of Health, Humboldt-University of Berlin, Berlin

Introduction

Telemedicine is increasingly used for long-term monitoring of chronically ill patients, such as kidney transplant recipients, but continuous streams of patient-reported data and messages impose a substantial workload on clinical staff. AI-supported communication tools offer the potential to reduce repetitive tasks while maintaining high-quality patient interaction.

Methods

We developed a chat-support system integrated into the SmartNTx telemedicine program at Charité [1]. The system generates suggested messages based on patient-reported vitals, abnormal values, and reminders for appointments or missing measurements [2]. All messages are reviewed by telemedicine staff before being sent. Initial templates were generated using LLMs and refined collaboratively with clinicians to ensure reliability and adherence to communication standards. The system has been in productive use since May 2024, covering 758 patients, and its performance was monitored until January 2026. Small-scale staff feedback was collected to evaluate usability, limitations, and perceived workload impact.

Results

During the observation period, >22k message suggestions were generated, of which >16k were sent, corresponding to an adoption rate of 76%. Most messages (95%) were sent without modification, indicating staff trust in the system. Telemedicine staff highlighted that reminders for missing measurements and appointments were most useful, whereas delays in message generation for rapidly changing vitals and redundant alerts were identified as key limitations. Staff noted potential for workload reduction, particularly for routine reminders, but stressed the importance of maintaining human oversight.

Discussion

Our experience demonstrates that AI-supported message suggestions can be successfully integrated into telemedicine workflows, supporting repetitive tasks and improving efficiency. However, deployment requires careful attention to organizational, infrastructural, and regulatory constraints. Hybrid approaches combining automated suggestions with clinician review are likely to be more acceptable in regulated healthcare settings than fully autonomous systems. Future work will explore scalability to additional centers and further integration with dynamic patient data.

References

[1] Schiffer, et al. The SmartNTx-study: a prospective, randomized controlled trial to investigate additional interventional telemedical management versus standard aftercare in kidney transplant recipients. Frontiers in Nephrology., 2025

[2] Duettmann et al. eHealth and Telemedicine for Risk Prediction and Monitoring in Kidney Transplantation Recipients. Technology in Healthcare, 2024

AI-based statistical tools with guardrails: building and validating a chatbot for sample size calculation

12:00-12:15

Presenter: Ewan Carr

Authors: Ewan Carr [1]; Sean Obi [2]; Oyebayo Ridwan Olaniran [1]; Diana Shamsutdinova [3]; Sarah Markham [1]; Daniel Stahl [1]; Gordon Forbes [4]

Affiliations: [1] King's College London, London; [2] Imperial College, London, London; [3] King's College London, LONDON; [4] King's College London, London

Introduction

Rigorous tools for sample size estimation in clinical prediction model (CPM) development exist, but uptake among applied researchers remains low. Key inputs such as outcome prevalence, expected predictive performance, and number of candidate predictors can be difficult to specify, and outputs require careful interpretation. Researchers increasingly use general-purpose large language model (LLM) chatbots for guidance, but these may hallucinate plausible-sounding yet unvalidated numerical estimates. What is missing is an interface that is both accessible and trustworthy.

Methods

We propose a framework for trustworthy LLM-based interfaces to statistical software with three components. First, knowledge-base grounding: factual claims are restricted to an expert-authored, literature-grounded domain knowledge base. Second, tool-calling for numerical output: all quantitative results are returned exclusively by established statistical software, with no mechanism for the LLM to generate numbers independently. Third, domain-expertise prompt engineering: best-practice guidance is encoded explicitly, including conservative choices when inputs are uncertain and correct interpretation of outputs. We demonstrate the framework in a web application based on pmsims, an R package for simulation-based CPM sample size estimation [1].

Results

The web application provides two interfaces: an interactive explorer for scenario visualisation and a guided chatbot for conversational elicitation. All sample size estimates returned by the chatbot are constrained to pre-computed estimates from pmsims, not generated by the LLM. We compare the accuracy of responses from a general-purpose LLM (GPT-4o) with no domain grounding against those from the chatbot across identical scenarios, and show how the framework extends to other statistical tools.

Discussion

This framework addresses both dimensions of the uptake problem, accessibility and correctness, without sacrificing methodological rigour. By keeping computation in established statistical software and interaction in the LLM, it opens these tools to a broader audience while preserving their statistical validity. While integrating LLMs into statistical workflows carries inherent risks, the guardrails presented here demonstrate that these can be mitigated by design. Applied broadly, this approach could increase uptake of best-practice methods across statistical software domains.

[1] Olaniran OR, Shamsutdinova D, Markham S, Zimmer F, Stahl D, Forbes G, Carr E. Adaptive Gaussian process search for simulation-based sample size estimation in clinical prediction models. https://arxiv.org/abs/26

From Plan to Code: A Randomized Coding Experiment on How Specificity of a Statistical Analysis Plan Shapes Reproducibility

12:15-12:30

Presenter: Daniela Dunkler

Authors: Daniela Dunkler [1]; Michael Kammer [2]; Sinan Necdet Cevirme [3]; Moritz Pamminger [1]; Georg Heinze [4]; Susanne Strohmaier [5]

Affiliations: [1] Medical University of Vienna, Center for Medical Data Science, Institute of Medical Statistics, Vienna; [2] Medical University of Vienna, Vienna; [3] Institute of Clinical Biometrics, Center for Medical Data Science, Medical University of Vienna, Vienna; [4] Medical University of Vienna, Center for Medical Data Science, Institute of Medical Statistics, Wien; [5] Department of Epidemiology, Center for Public Health, Medical University of Vienna, Vienna

Introduction

Recent meta-scientific advances highlighted that variation in scientific results arises from numerous, often undocumented choices across data preprocessing, modeling, and reporting, even when guided by a prespecified statistical analysis plan (SAP). To decrease the multiplicity of analyses strategies, preregistration of SAPs was proposed, but little attention was paid to variability introduced during the translation of SAPs into executable code.

Methods

We conducted a randomized coding experiment with 43 R coders being randomized to a standard SAP, as typically submitted in a preregistration, or a highly detailed SAP developed after thorough initial data analysis. All coders analyzed identical data and submitted code/outputs, and completed a questionnaire on their coding experience. One coder deployed AI agents to implement the standard SAP. A reference implementation by the study team served as benchmark.

Coders implemented a pharmacoepidemiologic case study using simulated data on individuals (n=125,000) hospitalized for acute coronary syndrome undergoing coronary stenting and treated with one of three P2Y12 inhibitors. The primary estimand was the 3-year restricted mean survival time (RMST).

The cohort size after preprocessing and the three RMST estimates were the main outcomes of the coding experiment to assess reproducibility.

Results

Implementations based on the detailed SAP yielded lower spread in cohort size (n from 18,179 to 22,146) and RMST estimates (e.g., P2Y12 inhibitor Prasugrel RMST from 864 to 1,018 days) with some submissions clustering around the benchmark (n=18,406 RMST 971 days). The standard SAP led to

wider variability (n from 6,894 to 19,968 RMST 760 to 973 days), reflecting heterogeneous interpretations. AI-assisted implementations exhibited variability comparable to human coders using the detailed SAP.

Discussion

Greater SAP specificity reduces implementation variability and improves reproducibility at the coding stage. However, achieving a sufficient level of detail to mitigate analysis variability requires in-depth initial data analysis and experience with the data, which may not be available at the stage of preregistration and entails trade-offs with practical feasibility. We discuss the different purposes SAPs can serve, like transparency of decisions, facilitating reproducibility, constraining analytic degrees of freedom, or protection against p-hacking, and their distinct requirements on the SAP.

[Project funded by ÖAW DATA_2023-32]

Semantic Health Data and Meta-Research

Mon, 28.09.26, 11:00–12:30, Room: N4

Chair: Danny Ammon

Extending the W3C Data Privacy Vocabulary (DPV) to cover the German Health Data Use Act (GDNG) - Lessons Learned

11:00-11:15

Presenter: Matthias Löbe

Authors: Julian Flake [1]; Christian Henke [2]; Matthias Löbe [3]; Alessandra Simone Kuntz [4]; Harshvardhan J. Pandit [5]; Christian Draeger [6]

Affiliations: [1] Institute for Software Technology (IST), University of Koblenz, Koblenz; [2] Universitätsmedizin Göttingen, Göttingen; [3] Universität Leipzig, Leipzig; [4] Universitätsmedizin Göttingen, Medizinische Informatik, Göttingen; [5] Trinity College Dublin, Dublin; [6] Universität Leipzig: Institut für Medizinische Informatik, Statistik und Epidemiologie, Leipzig

Introduction

Finding, sharing and reusing data requires the definition of a common vocabulary to adequately describe machine-actionable exchange data policies. The German Health Data Use Act (Gesundheitsdatennutzungsgesetz (GDNG)), which has been set into force in March 2024, regulates sharing and using health data for secondary use, such as research purposes and policy-making in Germany. It builds the German foundation for the future European Health Data Space (EHDS), a European initiative and regulation providing rules for sharing of health data within Europe.

Methods

Semantic Web vocabularies are a proven way to express such information describing data access and reuse. Descriptions get a stable and agreed meaning by formal definitions of terms that have a globally unique identifier. The Data Privacy Vocabulary (DPV) is such a vocabulary around privacy, data protection and related topics maintained by the W3C Data Privacy Vocabulary and Controls Community Group (DPVCG). DPV's main focus is on concepts extracted from data protection legislative texts, e.g. the European General Data Protection Regulation (GDPR) at the forefront. Additional extensions of the DPV are available to broaden the vocabulary's scope and applicability, such as extensions with terms extracted from EU's Data Governance Act or from EU's AI Act. The modular architecture of DPV allows for additional extensions, such as extensions for national regulations.

Results

In this article, we describe a new DPV draft extension dedicated to the German GDNG. It consists of 54 concepts extracted from the law. The concepts are applicable in different problem domains, where semantic web technologies are useful, for example to formalize data usage policies in ODRL.

The parallel development of the GDNG extension and an EHDS extension resulted in many shared concepts, which were contributed to another, overarching sector health extension.

Discussion

The GDNG vocabulary itself, the contributions to the sector health extension and the lessons learned from both may provide valuable insights for authors of similar vocabularies around the harmonization of EU member state regulations with the EHDS.

The full paper of this abstract will be published in the GMDS series "Studies in Health Technology and Informatics" with doi: 10.3233/SHTI261013, URL:https://dx.doi.org/10.3233/SHTI261013

A Qualitative Requirements Analysis for the Representation of Clinical Guidelines

11:15-11:30

Presenter: Bora Gashi, Fridtjof Schiefenhövel

Authors: Bora Gashi [1]; Fridtjof Schiefenhövel [2]; Martin Boeker [3]

Affiliations: [1] Technical University of Munich, School of Medicine, Klinikum rechts der Isar, Chair of Medical Informatics, Institute for Artificial Intelligence and Informatics in Medicine (AIIM), Munich; [2] Technical University of Munich, School of Medicine, Klinikum rechts der Isar, Department of Anaesthesiology and Intensive Care (AINS), Munich; [3] Technische Universität München (TUM), München

Introduction

Clinical Guidelines (CPGs, CGs) are mostly static, narrative documents, which may cause usability issues at the point of care. Although computable CG standards are available, major publishing portals do not publish them alongside CGs. This study aims to (1) provide a requirements analysis to formalize and visualize CGs, based on (2) a structured analysis of CGs sampled from 4 major CG portals, and (3) semi-structured qualitative interviews with clinicians from different medical specialties. The study seeks to identify a consolidated set of requirements to inform future efforts in CG formalization and design of CG-based CDSSs.

Methods

The structural document analysis included 20 full-version CGs obtained from AWMF (S3-level), NICE, SIGN and WHO. The extracted data covered structural features on the document and recommendation level, and were synthesized descriptively. Interviews were conducted via video call (25-35 min) with 8 physicians from the Klinikum Rechts der Isar (TUM), guided by a 14-question interview guide. Participants were purposefully recruited from a range of clinical specialties and seniority levels. Whisper (OpenAI) was used as a transcription tool, with subsequent human correction. Analysis was supported by Microsoft Copilot (Enterprise), via a structured prompt developed by the team. All outputs were reviewed by humans, with 3 quality checks applied.

Results

All 4 portals publish CGs as static narrative documents with no computable representations, spread across different documents. Interview analysis produced 45 categories in 5 thematic clusters, with 4 of them reported consistently: outdated or conflicting versions, insufficient specificity, inadequate or missing patient-specific applicability, and evidence gaps for rare diseases. Clinician challenges correspond directly to the structural analysis' findings. 22 requirements (functional, structural and visual) were derived.

Discussion

The 22 requirements based on 2 independent evidence sources provide a validated foundation for CG formalization and visualization.

The full paper of this abstract will be published in the GMDS series "Studies in Health Technology and Informatics" with doi: 10.3233/SHTI260993, URL:https://dx.doi.org/10.3233/SHTI260993

Mapping MMI Pharmindex Items to RxNorm Semantic Clinical Drugs - An Evaluation Study

11:30-11:45

Presenter: Markus Budeus

Authors: Markus Budeus [1]; Lucy Strüfing [2]; Editha Räuscher [3]; Martin Boeker [2]

Affiliations: [1] TUM University Hospital, Munich; [2] Technische Universität München (TUM), München; [3] TMF - Technologie- und Methodenplattform für die vernetzte medizinische Forschung e.V., Berlin

Introduction

Terminologies are essential for the interoperability of healthcare data and enabling high-quality medical research. The Medical Informatics Initiative Germany currently relies on data from the MMI Pharmindex to identify medications however, this database lacks international drug-level terminology. Therefore, this study aims to develop an algorithm to map RxNorm Semantic Clinical Drugs (SCDs) to medications in the MMI Pharmindex and to validate the accuracy of these mappings.

Methods

Using existing RxNorm identifiers at the substance level, we identified candidate RxNorm SCDs by using relationships within the RxNorm dataset. Candidates were retained if they matched MMI Pharmindex entries with respect to ingredients, strengths, and dose form. The resulting mappings were evaluated using a combination of automatic and manual validation procedures.

Results

At least one matching RxNorm SCD was identified for 18,326 medications. Validation indicated high precision and revealed opportunities to further improve coverage.

Discussion

While additional work is needed to improve RxNorm concept coverage, this study provides a strong foundation for coding medications in the Medical Informatics Initiative Germany using RxNorm, thereby facilitating the use of these data in international research.

The full paper of this abstract will be published in the GMDS series "Studies in Health Technology and Informatics" with doi: 10.3233/SHTI260979, URL:https://dx.doi.org/10.3233/SHTI260979

Semantic Retrieval of Projects Funded by the Innovation Committee

11:45-12:00

Presenter: Christopher Gundler

Authors: Christopher Gundler; Mareile Beernink; Michael Blaß; Alexander Johannes Wiederhold

Affiliations: Institute for Applied Medical Informatics, University Medical Center Hamburg-Eppendorf, Hamburg

Introduction

Evidence-based development of healthcare innovations requires efficient access to existing knowledge. The Innovation Committee of the Federal Joint Committee (Gemeinsamer Bundesausschuss, G-BA) has funded more than 800 projects since 2016, yet the heterogeneity of topics and terminology limits the effectiveness of the current keyword-based search interface.

Methods

To address vocabulary-mismatch problems and support more nuanced information needs, we developed and evaluated a semantic search tool using vector-based natural-language retrieval. The application embeds project descriptions and user queries and computes similarity directly in the browser, enabling privacy-preserving, offline operation. In a controlled comparative study with three experienced users, we assessed efficiency and effectiveness across four standardized search tasks.

Results

Although based on a small sample size, the results indicate that semantic retrieval can meaningfully improve the discovery of projects funded by the Innovation Committee and support evidence-based research workflows.

Discussion

Future work should validate these findings in larger cohorts and explore enhanced ranking strategies.

The full paper of this abstract will be published in the GMDS series "Studies in Health Technology and Informatics" with doi: 10.3233/SHTI261010, URL:https://dx.doi.org/10.3233/SHTI261010

Robustness checks as a practical tool to improve model building strategies - a paired Bayesian-frequentist experiment

12:00-12:15

Presenter: Michael Kammer

Authors: Michael Kammer [1]; Georg Heinze [1]; Aliaksandr Hubin [2]; Mariana Nold [3]

Affiliations: [1] Medical University of Vienna, Vienna; [2] Norwegian University for Life Sciences, Ås; [3] University of Jena, Jena

Introduction

For decades, statistical inference has been dominated by two paradigms: Bayesian and frequentist. In practice, the choice of either framework for data analysis is shaped by historical developments, computational complexity, research culture, and the analyst's training. However, despite their fundamental methodological differences, the two paradigms may yield comparable conclusions. We propose that when one of the two paradigms has been used for a statistical analysis, a robustness check under the other paradigm can help in uncovering non-critical and critical modeling assumptions by assessing similarities and differences of substantive results.

Methods

We conducted a paired-design experiment in four real-world case studies spanning prediction, diagnosis, and descriptive modeling. For each case, one researcher developed a statistical analysis plan (SAP) based on their preferred analytical framework, while a second researcher - favoring the other framework - developed a robustness check for the main analysis. The data was analyzed in parallel and the numerical results and conclusions were compared.

Results

We outline our proposed workflow by the example of diagnosing bacteria in the bloodstream from laboratory data in 4000 individuals. The estimated out-of-sample discrimination (c-index 0.76, 95% confidence interval [0.75, 0.80], and 0.76, 95% credible interval [0.74, 0.80], for frequentist and Bayesian analyses, respectively) and calibration (calibration slope 0.92 [0.78, 1.00] and 1.04 [0.88, 1.21]) were comparable between main analysis and robustness check. However, the two approaches differed substantially in model complexity and how they handled missing data during performance estimation. Similarly, for the other case studies, the substantive conclusions were consistent between Bayesian and frequentist analyses, but the paradigms diverged in the propagation of uncertainty, model selection and computational complexity.

Discussion

The Bayesian-frequentist divide, often portrayed as irreconcilable, appears to be context-dependent in practice. Embedding a cross-paradigm robustness check in a SAP subjects its statistical reasoning to thorough methodological scrutiny. Therefore, an agreement in the conclusions between the main analysis and the robustness check strengthens confidence in the substantive findings. Our case studies demonstrated that this integrative approach not only provides validation, but also fosters

collaboration to illuminate different aspects of a statistical problem that either paradigm alone might miss.

Multiverse-Style Analyses for Confirmatory, Exploratory and Meta-Research Purposes: Design, Conduct, and Interpretation

12:15-12:30

Presenter: Simon Lemster

Authors: Simon Lemster [1]; Alyssa Columbus [1, 2]; César Andrés Pérez Castro [1]; Daniela Dunkler [3]; Carsten Oliver Schmidt [4]; Ludwig Hothorn [5]; Sabine Hoffmann [1]; Anne-Laure Boulesteix [1]

Affiliations: [1] LMU Munich, Munich; [2] Johns Hopkins University Bloomberg School of Public Health, Baltimore; [3] Medical University of Vienna, Vienna; [4] Universität Greifswald, Greifswald; [5] Leibniz Universität Hannover, Hannover

Introduction

Multiverse-style analysis (MSA), such as vibration of effects analysis or specification curve analysis, is increasingly used in both meta-research and applied statistical settings, for example in epidemiology. However, there is little consensus on how MSA should be designed, conducted, or interpreted. We argue that while MSA can be performed in various ways, the concrete implementation should be guided by a clearly predefined research aim.

Methods

We propose three distinct workflows for conducting a MSA. Confirmatory MSA seeks to draw clear conclusions by performing inference across a small set of reasonable specifications targeting a precisely defined research question and estimand. Exploratory MSA likewise considers only reasonable specifications, but allows greater flexibility, for example with respect to estimands, modeling approaches, and visualizations. Exploratory MSA does not support inference its primary aims are robustness checks and hypothesis generation. Meta-research MSA has the overarching aim of characterizing and improving research processes, for example by reflecting the influence of different analytical choices in previously conducted studies. Using data from the Study of Health in Pomerania (SHIP), a large German population-based cohort study, we illustrate all three approaches for the research question of how body composition relates to cardiovascular mortality (CVM).

Results

The confirmatory MSA focuses on the causal effect of waist-to-hip ratio (WHR) on CVM, using a narrow multiverse of 24 specifications with different adjustment sets and outcome classifications. Accounting for multiplicity with the multiple marginal models approach, a significant effect of WHR on CVM is identified. The exploratory MSA considers a broader multiverse of 21,504 specifications, incorporating alternative exposure measures and covariate sets, and revealing substantial heterogeneity in results across specifications. The meta-research MSA with over 500,000 specifications

shows a non-negligible influence of both data preprocessing and model uncertainty, with analytical specifications derived from recent literature on body composition and cardiovascular disease.

Discussion

Our findings demonstrate how contradictions in the MSA literature can be addressed by clearly stating the purpose of the MSA and implementing it accordingly, following our recommendations. Even though distinctions between these approaches are not perfectly sharp in every context, these principles provide a practical framework towards enhancing research transparency and reliability.

Survival Inference

Mon, 28.09.26, 11:00–12:30, Room: Runder Saal

Chair: Marta Spreafico

Efficiency and interpretation of marginal versus conditional covariate adjustment for the Cox model

11:00-11:15

Presenter: Robin Ristl

Authors: Robin Ristl; Laetitia Fritz

Affiliations: Medical University of Vienna, Center for Medical Data Science, Institute of Medical Statistics, Wien

Introduction

Generalized covariate-adjustment in randomized trials is a field of active research and of increasing regulatory interest. These methods aim to increase the precision of a marginal (with respect to covariates) treatment effect estimator by explaining covariate-dependent variability. Recent publications show that marginal covariate adjustment applied to the Cox model hazard ratio may increase power compared to unadjusted analyses. However, due to non-collapsibility and the proportional hazards assumption, it is not obvious whether marginal adjustment or a usual conditionally covariate-adjusted Cox model is preferable. A further challenge with covariate-adjusted models pertains to the correct calculation of stopping-boundaries for interim analyses.

Methods

We considered trials with a survival endpoint under event-driven study design with stratified randomization and possible interim analyses. Hazard functions were assumed to depend on binary stratification factors and continuous covariates. We studied scenarios under proportional hazards conditional on covariates (PH), non-proportional hazards (NPH) and marginally proportional hazards (MPH). The latter was included as set-up where the true treatment effect corresponds to the target of marginal covariate adjustment. While the shape of hazard functions matching MPH in presence of covariate effects is not obvious, an according scenario was constructed by iteratively adjusting the hazard functions. Operating characteristics of unadjusted, conditionally and marginally adjusted Cox models and logrank tests were investigated by simulation.

Results

In the investigated PH scenarios, conditionally adjusted Cox models provided substantially more power than analogous marginally adjusted models. Under NPH, marginally adjusted tests were in general more powerful. In our MPH setting, the conditional adjustment was slightly more powerful. In all settings, either adjustment was more powerful than an unadjusted analysis. To complement the

inference procedures, we provide an algorithm for confidence intervals based on the inversion of (adjusted) logrank tests. We further propose an estimator for the covariance of group sequential statistics based on stacked estimating equations, to be used in the calculation of stopping-boundaries.

Discussion

Our results strengthen the use of conditionally adjusted Cox models as long as the proportional hazards assumption is at least approximately met. Under NPH, marginal adjustment of the hazard ratio provides increased power, however other effect measures may deserve attention in future comparisons.

Model-based bootstrap inference for Cox models after Lasso variable selection

11:15-11:30

Presenter: Lena Schemet

Authors: Lena Schemet [1]; Sarah Friedrich-Welz [2, 3]; Andreas Groll [4]

Affiliations: [1] University of Augsburg, Augsburg; [2] Institut für Mathematik, Universität Augsburg, Augsburg; [3] Center for Advanced Analytics and Predictive Sciences (CAAPS), Universität Augsburg, Augsburg; [4] Department of Statistics, TU Dortmund University, Dortmund

Introduction

Statistical inference after variable selection in high-dimensional survival analysis remains challenging. In Cox proportional hazards models, Lasso penalties are widely used for variable selection but lead to estimators with complicated sampling distributions, which hinders the construction of valid confidence intervals for regression parameters.

Methods

We investigate a model-based bootstrap approach for inference after Cox Lasso selection. The procedure generates bootstrap event times from the fitted Cox model obtained after penalized estimation. Specifically, the baseline cumulative hazard is estimated via the Breslow estimator conditional on the fitted regression coefficients. Bootstrap event times are then simulated using inverse transform sampling from the corresponding fitted survival distribution. To preserve the censoring mechanism, bootstrap censoring times are generated from an estimated censoring distribution. The Cox Lasso is subsequently refitted for each bootstrap sample to obtain bootstrap replicates of the regression coefficients. We further establish a first-order asymptotic justification of the proposed bootstrap procedure.

Results

Simulation studies show that the proposed bootstrap procedure provides reliable coverage of confidence intervals for selected regression coefficients across a range of sample sizes and sparsity levels typical for biomedical survival studies.

Discussion

The proposed model-based bootstrap offers a practical framework for post-selection inference in high-dimensional Cox regression. It preserves the semiparametric structure of the Cox model while enabling straightforward implementation using standard survival analysis software.

Model assessment with censored and truncated data: a new class of goodness-of-fit tests

11:30-11:45

Presenter: Jacobo de Uña-Álvarez

Authors: Juan Carlos Escanciano [1]; Jacobo de Uña-Álvarez [2]

Affiliations: [1] Universidad Carlos III de Madrid, Madrid; [2] Universidade de Vigo, Vigo

Introduction

Goodness-of-fit methods for censored and truncated data exist. However, they are typically based on the departure between the nonparametric maximum likelihood estimator and the fitted parametric model, and this may entail severe difficulties, such as nonexistence, non-uniqueness, computational instability, slow convergence rates or even violation of conditions for weak convergence.

Methods

We develop a systematic, omnibus approach to goodness-of-fit testing for parametric distributional models when the variable of interest is only partially observed due to censoring and/or truncation. For this, we construct a regular (pathwise differentiable) Neyman-orthogonal score process indexed by test functions, and aggregating it over a reproducing kernel Hilbert space ball. This yields a maximum-mean-discrepancy-type supremum statistic with a convenient quadratic-form representation. Critical values are obtained via a multiplier bootstrap that keeps nuisance estimates fixed.

Results

We establish asymptotic validity under the null and local alternatives and provide concrete constructions for left-truncated right-censored data, current status data, and random double truncation in particular, to the best of our knowledge, we give the first omnibus goodness-of-fit test for a parametric family under random double truncation in the composite-hypothesis case. Simulations demonstrate size control and power in practically relevant incomplete-data designs. We illustrate the proposed method by testing for a constant risk in disease free survival for bone marrow transplants under right-censoring. We also apply the method to testing for uniform truncation in the analysis of ages at diagnosis for childhood cancer with interval sampling, which is relevant to assess the potential observational bias.

Discussion

See Escanciano and de Uña-Álvarez (2026) (available at https://arxiv.org/abs/2602.08108) for further details.

References:

Escanciano JC, de Uña-Álvarez J (2026) Goodness-of-Fit Tests for Censored and Truncated Data: Maximum Mean Discrepancy Over Regular Functionals. arXiv:2602.08108.

Borrowing from historical control data in a Bayesian time-to-event model with flexible baseline hazard function

11:45-12:00

Presenter: Alex Lewin

Authors: Alex Lewin [1]; Darren Scott [2]

Affiliations: [1] London School of Hygiene and Tropical Medicine (LSHTM), London; [2] AstraZeneca, Cambridge

Introduction

Leveraging control-group data from historical trials is increasingly important for accelerating drug development. In time-to-event settings, existing Bayesian approaches often rely on restrictive assumptions about the baseline hazard to simplify computation, which can limit borrowing performance.

Methods

We propose a semi-parametric survival model using smoothing priors to estimate a fully flexible smooth baseline hazard function. Dynamic borrowing is achieved through a lump-and-smear prior on baseline hazard parameters, allowing the model to learn adaptively from historical data. Our work includes a principled strategy for setting values of hyperparameters to control the level of dynamic borrowing, based on the tolerated difference between current and historical baseline hazards.

Results

We demonstrate the benefit of the dynamic borrowing prior through a simulation study. We show improved power when historical data are exchangeable with current data, whilst reducing type I error in the presence of prior-data conflict. We further illustrate the method using data splitting on a trial for a treatment used in asthma.

Discussion

A flexible approach to modelling the baseline hazard allows for dynamic borrowing in trials involving survival and time-to-event outcomes.

Statistical Inference for Time-to-Event Outcomes using Survival Data Stacking

12:00-12:15

This is one of this year's ISCB CASP award-winning abstracts.

Presenter: Hui Wen

Authors: Hui Wen; Ionut Bebu; Guoqing Diao

Affiliations: George Washington University, Washington, DC

Introduction

Standard modeling approaches for event-time outcomes include parametric (e.g., Weibull) and semi-parametric (e.g., Cox proportional hazards [PH]) models. Machine learning (ML) approaches (e.g., random survival forests) have also been considered but are not straightforward to extend to settings such as time-varying exposures or recurrent events. Recently, event-time analysis has been reframed as a sequence of classification problems using survival data stacking, enabling flexible ML methods. However, its performance in estimating individual survival probability and restricted mean survival time (RMST), particularly under non-PH settings, remains unclear.

Methods

We conducted simulation studies comparing survival stacking with a range of machine learning approaches for estimating survival curves and RMST under proportional and non-proportional hazards settings. To enable statistical inference, we derived delta method estimators for survival probabilities and RMST within a logistic regression framework applied to the stacked data. Corresponding RMST inference from Cox models was used as a benchmark for comparison.

Results

Across simulation settings, logistic regression within the survival stacking framework achieved performance comparable to Cox models under proportional hazards and demonstrated improved accuracy under non-proportional hazards. The proposed inference approach yielded coverage probabilities close to nominal levels across time horizons for both survival estimates and RMST.

Discussion

These findings highlight survival stacking as a flexible and robust framework for estimating clinically relevant quantities under complex hazard structures, particularly in settings where standard model assumptions may not hold. Logistic regression within the stacked framework yields stable estimates and enables interpretable uncertainty quantification for both survival probabilities and RMST. Ongoing work extends this approach using M-estimation to improve robustness beyond model assumptions. The approach is further illustrated through an application to cardiovascular disease risk in individuals with type 1 diabetes from the DCCT/EDIC study.

Identifying Hidden Mortality Phenotypes in Net Survival: A Latent Class Approach to Excess Mortality

12:15-12:30

Presenter: Anouar Abdeldjaoued Ferfache

Authors: Anouar Abdeldjaoued Ferfache [1]; Nathalie Graffeo [1]; Roch Giorgi [2]

Affiliations: [1] Aix Marseille Univ, Inserm, IRD, SESSTIM, Sciences Economiques & Sociales de la Santé & Traitement de l'Information Médicale, Equipe Labellisée Ligue Contre le Cancer, ISSPAM, Marseille, France, Marseille; [2] Aix Marseille Univ, APHM, Inserm, IRD, SESSTIM, Sciences Economiques & Sociales de la Santé & Traitement de l'Information Médicale, Equipe Labellisée Ligue Contre le Cancer, ISSPAM, Hop Sainte Marguerite, BioSTIC, Marseille, France, Marseille

Introduction

Net survival analysis is one the main approach in population-based cancer survival research. However, standard excess hazard regression models often overlook the inherent heterogeneity within patient populations. Indeed, they typically assume that covariate effects on excess mortality are uniform, potentially masking distinct subgroups with divergent risk profiles. This study introduces a Latent Class Excess Mortality Model (LCEMM) to identify unobserved 'mortality phenotypes' within the net survival framework.

Methods

We developed an LCEMM that allows for the simultaneous estimation of class-specific hazards and membership probabilities. To ensure the statistical integrity of the class-specific intercepts, we incorporated a Baseline-Corrected Delta Method. This accounts for the complex interdependence between the the class-specific intercepts and the normalized baseline excess hazard. The methodology was validated through a comprehensive simulation study designed to assess the model's ability to recover distinct latent risk groups under varying degrees of class overlap and censoring.

Results

The simulation study confirmed that the latent class framework successfully identifies 'hidden' high-excess-hazard subgroups that are otherwise obscured in aggregate net survival models. Methodologically, we found that traditional inference techniques significantly underestimate the uncertainty of these latent classes. In contrast, our proposed normalization-corrected approach restored standard errors to their true empirical levels, ensuring reliable confidence intervals and nominal coverage.

Discussion

This research provides a novel tool for public health researchers to move beyond aggregate survival metrics. By identifying latent risk classes in net survival, we can better characterize 'hidden' populations facing distinct survival trajectories. This approach provides a mathematically rigorous foundation for more personalized population-based surveillance and a deeper understanding of health disparities in cancer outcomes.

13:30–15:00

LLMs for Clinical Workflows

Mon, 28.09.26, 13:30–15:00, Room: K2

Chair: Fabian Kabus

Rule-Based Prompting for Extraction of TNM Stage and Tumor Grade from Population-Level Lithuanian Pathology Reports Using a Medical Large Language Model

13:30-13:45

Presenter: Tadas Žvirblis

Authors: Tadas Žvirblis [1]; Tomas Griguola [2]; Justas Mileika [1]; Paulius Milmantas [1, 3]; Miglė Gervytė [1, 4]; Agnė Irkinaitė [1]; Valdas Pečeliūnas [1]

Affiliations: [1] Cancer Registry, Laboratory of Cancer Epidemiology, National Cancer Institute, Vilnius; [2] National Cancer Institute, Vilnius; [3] Institute of Informatics, Faculty of Mathematics and Informatics, Vilnius University, Vilnius; [4] Institute of Data Science and Digital Technologies, Faculty of Mathematics and Informatics, Vilnius University, Vilnius

Introduction

Pathology reports contain key oncological variables, including tumor grade and TNM stage, but these are typically embedded in unstructured free text. This limits secondary use for cancer surveillance, outcomes research, and clinical data integration, particularly in low-resource languages such as Lithuanian. Large language models (LLMs) may provide a scalable approach for structured information extraction from routine clinical text.

Methods

We evaluated MedGemma-27B for extraction of tumor grade (G) and clinical (c) / pathological (p) TNM variables (cT, cN, cM, pT, pN, pM) from Lithuanian pathology and related clinical reports collected since 2020 according to TNM 8th edition definitions. The dataset represented routine population-level real-world data, with no preselection by tumor site, diagnosis, or other clinical criteria. Two prompting strategies were compared: basic JSON prompt and iteratively refined rule-based prompt with explicit instructions for distinguishing clinical from pathological staging and assigning null values when evidence was absent or ambiguous. Prompts were tested in Lithuanian and English. Performance was evaluated on 100,000 manually annotated reports using accuracy and extraction coverage.

Results

Rule-based prompting improved extraction accuracy for most variables, including tumor grade (G: 0.83 vs 0.70) and clinical staging components (cN: 0.43 vs 0.32 cM: 0.49 vs 0.42). Accuracy also improved for pathological variables (pT: 0.79 vs 0.73 pN: 0.70 vs 0.66). Accuracy for pM declined under structured prompting (0.44 vs 0.55). Clinical T-stage extraction remained challenging under both prompting strategies (cT: 0.38 vs 0.27). Prompt language had no meaningful effect on performance. However, coverage varied across variables, with increases observed for some (e.g., G, pT) and decreases for others, particularly for rare variables such as pM.

Discussion

In heterogeneous, non-selected real-world pathology data, prompt design materially affected both accuracy and completeness of LLM-based extraction. Rule-based prompting improved accuracy across most variables without model retraining, supporting the feasibility of structured oncology data extraction in a low-resource language setting. However, these improvements were accompanied by variable changes in extraction coverage, with reduced completeness observed for some variables. Persistently low accuracy for cT and reduced performance for rare variables such as pM highlight limitations of prompt-based approaches. Task-specific validation remains essential before downstream clinical or epidemiological use.

LLM-Based Semi-Automated Information Extraction from Histopathological Reports in T-Cell Lymphoma Patients

13:45-14:00

Presenter: Christoph Demus

Authors: Christoph Demus [1]; Filip Kramarzyk [2]; Annemarie Weise [1]; Christian Jäger [1]; Thomas Weber [2]; Jan Christoph [1]

Affiliations: [1] Junior Research Group (Bio-)Medical Data Science, Faculty of Medicine, Martin-Luther-University Halle-Wittenberg, Halle (Saale); [2] Department of Internal Medicine IV, Hematology and Oncology, University Hospital Halle (Saale), Halle (Saale)

Introduction

Clinical registries are essential for research on rare and heterogeneous diseases such as T-cell lymphomas. However, manual extraction of relevant information from unstructured clinical documents remains time-consuming and limits data completeness and quality. Large language models (LLMs) offer the opportunity to support information extraction from clinical routine data. The DQ-LIR project aims to improve data density and quality in entity-specific clinical registries by implementing an LLM-based pipeline for semi-automated extraction of structured information from histopathology reports.

Methods

We developed a locally hosted LLM-based processing pipeline to extract predefined clinical variables from pseudonymized histopathology reports of patients with T-cell lymphomas. The training dataset contained 218 reports from 68 patients recorded in the University Hospital Halle. After text extraction and preprocessing of PDF reports, relevant variables (including diagnosis, tissue type, ICD-10 code, and multiple immunohistochemical markers) were automatically extracted using LLMs with specific prompts for each marker. The system combines prompt-based LLM extraction with rule-based preprocessing, including regular-expression filtering of relevant text passages and synonym mappings. Performance was evaluated against a manually annotated gold standard created by medical experts. To assess the reliability of the annotators, the inter-annotator agreement was also determined. Three open-source LLMs were compared, and iterative optimization was performed through repeated evaluation cycles.

Results

Among the tested models, Llama-3.1-8B-Instruct achieved the best overall performance with a mean accuracy of 0.90. Extraction accuracy varied across variables: several immunohistochemical markers achieved accuracies above 0.95, while more complex tasks such as diagnosis classification showed lower performance (accuracy of 0.84). The integration of rule-based preprocessing significantly improved results, primarily by reducing hallucinations when information was absent in the reports. Extracted data were exported into registry-compatible CSV files for semi-automated integration into the clinical registry.

Discussion

Our results demonstrate that LLM-based approaches can substantially support the extraction of structured registry data from unstructured pathology reports. A hybrid architecture combining LLMs with rule-based preprocessing proved particularly effective in this specialized medical domain. The locally deployed system enables privacy preserving use in a clinical environment and has the potential to reduce manual workload while improving registry data completeness.

From Free-Text to ICD-10 Codes: Comparison of Embedding Models and Retrieval-Augmented Generation for Automated Coding of German Tumor Diagnoses

14:00-14:15

Presenter: Fatma Alickovic

Authors: Fatma Alickovic [1]; Stefan Lenz [1]; Arsenij Ustjanzew [1]; Lakisha Ortiz Rosario [1]; Georg Vollmar [2]; Thomas Kindler [2]; Torsten Panholzer [1]

Affiliations: [1] Institute of Medical Biostatistics, Epidemiology and Informatics, University Medical Center of the Johannes Gutenberg University Mainz, Mainz; [2] University Cancer Center of the Johannes Gutenberg University Mainz, Mainz

Introduction

Accurate ICD-10 coding of free-text tumor diagnoses is essential for cancer registries and research. However, the heterogeneity of diagnostic text makes automated coding difficult. While retrieval-augmented generation (RAG) [1] and embedding models [2] have been explored for this task, the key objective in this study is to assess the added value of LLMs and RAG over purely embedding-based approaches.

Methods

Our evaluation is based on a real-world test dataset from the local tumor documentation system that contains short diagnosis descriptions and corresponding codes (2,024 entries 238 distinct codes). The Alpha-ID (12,021 entries 776 codes) and ICD-10-GM classification (859 entries 859 codes) served as a knowledge base. Nine different embedding models were used to predict codes via semantic search and alternatively via an added neural network classifier trained on the knowledge base embeddings. For RAG, the top-3 retrieved entries were incorporated into prompts for Llama-3.1-8B and Llama-3.3-70B to generate ICD-10 codes. The performance was evaluated using accuracy of exact-match (full codes) and partial-match (three-characters only) of predicted codes in the test dataset. A data quality analysis indicated a maximum of approximately 70% and 89% for attainable exact and partial-match accuracy.

Results

Among the embedding models, Qwen3-Embedding-8B achieved the best performance in both semantic search (44.3% exact-match, 67.7% partial-match accuracy) and combined with the classification network (47.8% exact-match, 72.1% partial-match accuracy). RAG substantially improved the performance of the base LLMs. RAG with Llama-3.3-70B achieved the highest partial-match accuracy (81%) but did not surpass Qwen3-Embedding-8B in exact-match performance, while Llama-3.1-8B did not outperform embedding approaches in either metric.

Discussion

Our results suggest that while RAG can enhance the coding performance of LLMs without requiring additional model training, LLMs with and without RAG are not necessarily better than pure embedding-based approaches. Fine-tuning embedding models could be a promising direction for future work.

[1] Krumscheid et al. (2025), Retrieval-Augmented Generation for ICD-10 Coding in German Clinical Texts - A Technical Case Report. https://doi.org/10.3233/SHTI251397

[2] Kreuzthaler et al. (2025), Secondary Use of Clinical Problem List Descriptions for Bi-Encoder Based ICD-10 Classification. https://pmc.ncbi.nlm.nih.gov/articles/PMC12099355/

Benchmarking AI Vibe Coding for Clinical Statistical Analysis: A Structured Evaluation in Pulmonary Hypertension Research

14:15-14:30

Presenter: Jeeva Sam

Authors: Jeeva Sam [1, 2]; Til Karsten Spreuer [1, 2]; Max Breger [1, 2]; Athiththan yogeswaran [1, 2]; Khodr Tello [1, 2]; Patrick Janetzko [1, 2]; Werner Seeger [1, 2]; Jonas Bienzeisler [3]; Raphael W. Majeed [1, 2, 3]

Affiliations: [1] Department of Internal Medicine, Universities of Giessen and Marburg Lung Center (UGMLC), Member of the German Center for Lung Research (DZL), Gießen; [2] Cardio-Pulmonary Institute (CPI), Institute for Lung Health (ILH), Gießen; [3] Medizinische Faculty, RWTH University, Aachen, Aachen

Introduction

Statistical analysis of clinical data requires expertise in medical statistics. Large language models (LLMs) are increasingly used for code generation and may support both descriptive and advanced analyses, but their reliability remains uncertain. Errors in grouping, missing-value handling, or model specification may produce outputs that appear plausible while remaining analytically incorrect. This study benchmarked the reliability of current LLMs for predefined clinical statistical analyses against a verified expert reference.

Methods

Five LLMs (GPT 5.3, Claude Sonnet 4.6, Gemini 2.5 Flash, Perplexity, and Grok) were evaluated across three statistical tasks derived from a pulmonary hypertension registry workflow: descriptive table generation, Kaplan-Meier survival analysis, and Cox proportional hazards modelling. All models received identical datasets and standardised prompts. Outputs were compared with analyses performed by two experts trained in mathematical statistics and assessed for grouping correctness, numerical accuracy, missing-value handling, completeness of requested output, and quality of generated R code. Benchmark datasets were derived from the PVRI GoDeep meta-registry and reflected the structure of a published pulmonary hypertension analysis workflow.

Results

All five models produced correct descriptive statistics once dataset variables were specified explicitly. In the initial descriptive benchmark, two models failed because of variable-name ambiguity, but both recovered after prompt clarification. All models reproduced the correct Kaplan-Meier log-rank p-value, although figure completeness differed. In the Cox benchmark, four models reproduced all required hazard-ratio terms, while one omitted the interaction terms required for the full analysis. Even when statistical outputs were correct, substantial differences remained in code quality, formatting, and direct usability.

Discussion

Current LLMs can support routine clinical statistical tasks, but their reliability depends strongly on prompt precision and careful verification of generated code and results. Numerical correctness alone is not sufficient to assess practical usefulness, as code quality and completeness of output varied substantially between models. LLMs may therefore be useful as assistive coding tools in clinical research workflows, but they should not be regarded as autonomous analytical systems. Independent validation remains necessary before generated outputs are used in research practice.

The full paper of this abstract will be published in the GMDS series "Studies in Health Technology and Informatics" with doi: 10.3233/SHTI261007, URL:https://dx.doi.org/10.3233/SHTI261007

LLM-Assisted Generation of Off-Label Medication Requests in Oncology: A QUEST-Based Evaluation Framework

14:30-14:45

Presenter: Valeriya Vishnevskaya

Authors: Valeriya Vishnevskaya [1]; Georg Mathes [1]; Stephanie Berger [2]; Stefan Sigle [1]

Affiliations: [1] MOLIT Institute gGmbH, Heilbronn; [2] SLK Kliniken Heilbronn, Heilbronn

Introduction

Off-label medication requests in precision oncology require extensive documentation of a patient´s treatment history, molecular diagnostics and exhaustion of guideline-based therapies. Generative approaches like Large language models (LLMs) bear potential supporting this workflow but their evaluation remains challenging due to scarcity of domain experts and the nature of generative approaches [1].

Methods

In literature the QUEST framework has been identified as a suitable framework evaluating principles like Quality, Understanding/Reasoning, Expression, Safety and Trust of LLMs in healthcare [2]. Based on QUEST we propose a layered evaluation scheme: Layer 1 (a), applicable without clinical expertise, scores on four dimensions rated on a 1-5 Likert scale: factual faithfulness to input data, patient case completeness, appropriate wording, and clinical readiness. A binary red-flag checklist covers hallucinations, incorrect cancer staging, omitted molecular findings or treatment history, and unsupported guideline citations. Layer 2 (b) extends this process by open-ended questions for domain experts, e.g. did the given output correctly identify which standard options had been used? A joint evaluation of (a) and (b) should give an indication on how a LLM based system for Off-Label requests performs.

Results

Preliminary testing of the proposed methodology identified three failure categories: (1) inference beyond source data: e.g. interpreting lung biopsy as evidence of operability in stage IV patients (2) mismatch between generated medication recommendations and regulatory approval status and (3) hallucination combined with unjustified certainty under data incompleteness.

Discussion

Primary risk factors are hallucination combined with unjustified arguments for personalized off label generation. The proposed framework aims to provide pragmatic evaluation focused on safety and usability. It should uncover safety-relevant failure modes prior to deployment and offers a transferable model for LLM evaluation in clinical documentation.

[1] J. Zhao, H. Liu, Y. Chen, and F. Song, "Application of artificial intelligence tools and clinical documentation burden: a systematic review and meta-analysis," BMC Med. Inform. Decis. Mak., vol. 26, no. 1, p. 29, Dec. 2025, doi: 10.1186/s12911-025-03324-w.

[2] T. Y. C. Tam et al., "A framework for human evaluation of large language models in healthcare derived from literature review," Npj Digit. Med., vol. 7, no. 1, p. 258, Sep. 2024, doi: 10.1038/s41746-024-01258-7.

IRIS: A Multi-Agent Large Language Model Framework for Conversational Clinical Trial Navigation and Protocol Design

14:45-15:00

Presenter: Paolo Messina

Authors: Amirmahdi Abootalebi; Paolo Messina; Mario Torchia; Luca Emili

Affiliations: InSilicoTrials Technologies Spa, Trieste

Introduction

Recent advances in large language models (LLMs) have demonstrated strong capabilities in biomedical natural language understanding, structured reasoning, and multi-turn dialogue, opening new possibilities for intelligent, context-aware clinical research support. However, the integration of LLMs into end-to-end clinical trial workflows remains an open area of investigation.

Objectives

We present the clinical trial reasoning component of IRIS, a multi-agent AI assistant designed to enable conversational, context-aware interaction with clinical trial registries. The primary objective is to demonstrate how an orchestrated pipeline of LLM-based agents can bridge the gap between unstructured user intent and structured clinical trial knowledge, supporting tasks such as eligibility criteria design, trial landscape exploration, and protocol comparison through natural language dialogue.

Methods

IRIS employs a pipeline of specialized, cooperating LLM-based agents. A query interpretation agent converts natural language prompts into structured search parameters, which a retrieval agent uses to access the ClinicalTrials.gov API. To ensure only the most semantically relevant records are passed downstream, a ColBERT-based reranking model scores and re-orders the retrieved candidates. A reasoning agent processes this set using a pre-trained LLM to synthesize design recommendations. Finally, a conversational agent maintains session context across multi-turn exchanges, enabling users to iteratively refine their queries, request comparisons, or explore alternative trial designs. In this work we present some results from the reranking model.

Results

Among the 10 documents retrieved per query, we evaluated the top 5 reranked, ColBERT outperformed the BGE (BAAI General Embedding) baseline across both metrics (mean Normalized Discounted Cumulative Gain @5: 0.970 vs. 0.927 Recall@5: 90% vs. 80%). The performance gap was most pronounced on semantically complex queries, where ColBERT scored 0.954 versus BGE's 0.866, confirming its advantage precisely in the contextually rich query types characteristic of clinical trial design tasks.

Discussion

IRIS demonstrates that a multi-agent LLM architecture, grounded in real-time clinical trial registry data, can meaningfully augment the early phases of trial design and literature navigation. This approach accelerates protocol ideation by surfacing design precedent through natural language. Future work will extend IRIS to additional registry sources, integrate structured outcome data for comparative effectiveness queries, and evaluate system-generated criteria against expert-reviewed gold standards.

Learning Latent Clinical Trajectories

Mon, 28.09.26, 13:30–15:00, Room: K3

Neural Dynamics-Enhanced Linear Mixed Models for Irregular Longitudinal Data

13:30-13:45

Presenter: Zhe Li

Authors: Zhe Li [1, 2]; Quentin Clairon [1, 2, 3]; Cécilia Samieri [1]; Mélanie Prague [1, 2, 3]; Cécile Proust-Lima [1]

Affiliations: [1] Université de Bordeaux, Inserm, Bordeaux Population Health Research Center, U1219, Bordeaux; [2] Inria, SISTM, Bordeaux; [3] VRI, Paris

Introduction

Longitudinal cohort studies generate irregularly spaced, partially missing repeated measurements where outcomes may depend on the full trajectory of time-varying exposures rather than their instantaneous values. Classical linear mixed-effects models (LMMs) are structurally limited to pre-specified and instantaneous parametric covariate effects, making them unable to capture path-dependent or cumulative exposure mechanisms. We propose the Neural Dynamics LMMs (ND-LMMs), a unified framework embedding continuous-time latent dynamics within a mixed-effects model with the Neural ODE/CDE as special cases differing in their sensitivity to the exogenous covariate path.

Methods

Baseline and static covariates are encoded into an initial latent state, which evolves over time via either a Neural ODE with autonomous dynamics or a Neural CDE where time-varying covariate change explicitly drive the trajectory. The latent process is decoded through a LMMs observation model, enabling estimation of population-level fixed effects and subject-specific random effects, with parameters learned by maximizing the marginal likelihood. To interpret covariate effects, we compute Partial Dependence Plots (PDPs) under counterfactual interventions and define the ΔPDP as the difference in population-averaged predictions between exposure levels with uncertainty quantified via a full-parameter delta method. The approach is evaluated in the simulation studies under diverse data-generating processes, and compared to the classical LMMs. It is also applied on the real population-based aging Trois-Cités (3C) cohort (N=7,324) to assess the associaion between cardiometabolic health (e.g., blood pressure, glycemia) and cognitive decline over 15 years follow-up.

Results

Under an instantaneous association (X(t) –> Y(t)), our approach achieves ΔPDP estimation performance comparable to the correctly specified classical LMM. Under a cumulative burden process, the

ND-LMMs correctly distinguishes sustained early exposure from a late spike to the same level, a discrimination the classical LMMs structurally cannot achieve. On the 3C cohort, our method delivers interpretable marginal associations between trajectories of cardiometabolic factors and cognitive decline via the PDP.

Discussion

The ND-LMMs extends classical LMMs with continuous-time neural dynamics while preserving their statistical interpretability. Simulation and real-data results demonstrate that the framework matches LMMs performance when covariate effects are instantaneous and outperforms it when effects are path-dependent or cumulative, a setting highly relevant in life-course epidemiology.

Multi-Outcome Disease Progression: Transformers and Deep Learning for Temporal correlation in Longitudinal Data

13:45-14:00

Presenter: Diana Shamsutdinova

Authors: Diana Shamsutdinova [1]; Sun Bin Kim [2]; Daniel Stahl [3]; Daniel Stamate [4]

Affiliations: [1] King's College London, LONDON; [2] King's College London, London; [3] King's College London, London; [4] Goldsmiths, University of London, London

Introduction

Multimorbidity, co-occurrence of multiple chronic diseases, is increasingly prevalent yet understudied in predictive modelling. Longitudinal observational datasets contain rich clinical information and are excellent sources for multi-outcome prediction. Causal prediction modelling must also be advanced, enabling models to provide clinical insights besides black-box individual risk estimates. We aim to 1) investigate a structured methodological trajectory, Cox → deep Cox → discrete-time neural models → Transformer architectures, to determine how increasing model complexity should match underlying data complexity and 2) advance causal prediction modelling by developing models capable of uncovering inter-temporal dependencies and cross-diagnosis patterns.

Methods

We designed a controlled, simulation-based framework to quantify the strengths of each modelling stage under known data-generating mechanisms. We extended classical Cox proportional hazards models to non-linear, multi-outcome deep Cox networks. Building on the Cox-Poisson likelihood equivalence, we transformed survival outcomes into discrete-time binary sequences, enabling flexible neural architectures to model multi-outcome interactions. We then reframe multimorbidity as a multivariate time-series problem. Existing transformer-based survival models (Delphi-2m, SurvTrace, SurvFormer) focus primarily on predictive performance here, we introduce a novel Transformer architecture prioritising interpretability of inter-temporal and cross-diagnosis patterns to support causal prediction modelling.

Our simulated populations with adjustable non-linearities, interactions, and temporal dependencies, allowed examining how model performance depends on alignment between model flexibility and data complexity. We will then apply the transformer model to the English Longitudinal Study of Ageing (ELSA), predicting multiple physical and mental health outcomes from socio-economic, clinical and diagnostic-history variables.

Results

In simulations with five correlated outcomes, deep models improved concordance by up to 20% over linear Cox when non-linearities or cross-terms were present. Our interpretability-oriented Transformer successfully recovered true inter-temporal dependencies and improved prediction in settings with substantial temporal structure. In ELSA (N≈10,000), gains are modest, suggesting that transformer-level benefits require either strong underlying dependencies or substantially larger sample sizes.

Discussion

Increasingly flexible models enhance predictive accuracy and interpretability when disease dynamics are complex. Transformers can reveal mechanistic temporal pathways relevant for causal prediction, but simpler survival models remain competitive in typical cohort settings. Model choice should therefore be guided by the complexity of the underlying data-generating process.

From Latent States to Clinical Insight: Interpreting Disease Trajectories Across Comorbidities

14:00-14:15

Presenter: Alice Flamigni

Authors: Alice Flamigni [1, 2]; Nicole Fontana [1, 2]; Francesca Ieva [2, 1]; Nicole Soranzo [1, 3, 4]

Affiliations: [1] Human Technopole, Milan; [2] MOX lab, Department of Mathematics, Politecnico di Milano, Milan; [3] Wellcome Sanger Institute, Hinxton; [4] University of Cambridge, Cambridge

Introduction

Modeling the longitudinal dynamics of chronic diseases in the presence of multiple comorbidities remains a major challenge, particularly when aiming to retain clinical interpretability. Many existing machine learning approaches capture complex patterns in patient trajectories, but provide limited insight into their clinical meaning. In this work, we introduce a unified framework that combines probabilistic modeling of disease progression with language-based interpretation to address both aspects simultaneously.

Methods

We employ an unsupervised Hidden Markov Model to infer latent disease states from longitudinal electronic health record data. The model is designed to integrate heterogeneous sources of information, including clinical variables and treatment data, while explicitly capturing the influence of genomic features on state transition dynamics. Beyond state inference, we incorporate a large language model to interpret the learned latent representations. The model systematically translates the statistical characteristics of each state into clinically meaningful descriptions, leveraging biomedical knowledge to contextualize patterns that would otherwise remain non interpretable.

Results

Our aim is to show that the inferred trajectories are consistent with established clinical knowledge while also revealing previously unrecognized progression patterns across comorbid conditions. Importantly, the integration of the large language model provides an additional interpretative layer, enabling a more transparent understanding of the latent states and their transitions, and facilitating their potential use in clinical decision-making. This combined approach highlights subtle disease dynamics and treatment effects that may not be fully captured by traditional analyses or expert-driven frameworks

Discussion

Overall, this work proposes an interpretable and extensible methodology for studying the longitudinal dynamics of comorbidities, bridging the gap between data-driven modeling and clinical insight.

Latent Class Mixed Model Variational Autoencoders for Scalable Phenotyping of Long-term Health Trajectories

14:15-14:30

Presenter: Philomène Letzelter

Authors: Philomène Letzelter [1]; Sandrine Katsahian [2]; Benjamin Vittrant [3]; Agathe Guilloux [4]

Affiliations: [1] Withings, Issy-les-Moulineaux, France, Paris; [2] URC HEGP, CIC-EC1418, AP-HP, Paris; [3] Withings, Issy-les-Moulineaux; [4] Inserm, Université Paris Cité, Inria, UMR 1346, HeKA, Paris

Introduction

Large-scale longitudinal datasets, such as Electronic Health Records and wearable data, offer unprecedented opportunities for retrospective epidemiological studies. However, they often contain dozens of correlated, irregularly sampled biomarkers, posing analytical challenges. Traditional Mixed-Effects Models provide interpretable inference for repeated measurements but scale poorly to high feature dimensions. Conversely, Variational Autoencoders (VAEs) offer flexible dimensionality reduction but inherently assume independent observations, failing to capture grouped longitudinal

structure. The LMM-VAE (Ong et al., 2025) bridges this gap with a structured linear mixed model (LMM) prior, enabling scalable and interpretable trajectory modeling, yet assumes a homogeneous cohort, leaving population heterogeneity unaddressed. No framework simultaneously enables scalable, high-dimensional longitudinal modeling and trajectory-grounded latent subgroup discovery.

Methods

We propose the Latent Class Mixed Model Variational Autoencoder (LCMM-VAE), extending LMM-VAE by replacing its global LMM prior with a Gaussian Mixture of class-specific LMMs. The prior is partitioned into: (1) population-level fixed effects shared across all classes, (2) class-specific fixed effects capturing divergent mean trajectories, and (3) class-specific random effects modeling individual deviations around the class trajectory. Each observation is encoded independently the class-conditional LMM prior imposes temporal and subgroup structure and derives per-subject cluster responsibilities. Training maximizes a variational lower bound over both continuous latent states and discrete class assignments, extending Ong et al.'s global parametrization to the mixture setting while preserving interpretability and compatibility with mini-batch training on large datasets.

Results

Evaluations on benchmark simulations and real-world longitudinal data demonstrate that LCMM-VAE recoversdistinct phenotypes based on trajectory dynamics rather than static baselines, with improved latent cluster separation relative to standard VAEs. Furthermore, the structured prior enables reconstruction of long-term biomarker trajectories from partial observations, without prior knowledge of an individual's progression stage.

Discussion

LCMM-VAE integrates classical latent class model interpretability with deep generative modeling flexibility, identifying unobserved phenotypes in high-dimensional longitudinal data. By disentangling shared trends, subgroup trajectories, and individual variations within a unified probabilistic framework, this method offers a promising tool for precision medicine and retrospective disease phenotyping.

References

Ong, P. et al. (2025). Latent mixed-effect models for high-dimensional longitudinal data. TMLR.

https://openreview.net/forum?id=7A96yteeF9

Synthetic Clinical Supervision with Large Language Models in Latent Disease Trajectories

14:30-14:45

Presenter: Clemens Schächter

Authors: Clemens Schächter [1]; Astrid Pechmann [2]; Janbernd Kirschner [2]; Harald Binder [1]

Affiliations: [1] Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg im Breisgau; [2] Department of Neuropediatrics and Muscle Disorders, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg im Breisgau

Introduction

Statistical and generative models can introduce artifacts due to model misspecification. In generative settings, these artifacts may result in synthetic data that yield a different clinical interpretation than the original observations. Correcting such inconsistencies through human expert supervision is often impractical due to cost and time constraints. Large Language Models (LLMs), which encode broad medical knowledge, offer a scalable alternative as synthetic clinical experts.

In this work, we integrate LLM-derived clinical judgement into the training of a variational autoencoder (VAE), a generative model used to learn low-dimensional representations of data. We investigate whether LLM-based supervision can reduce clinically implausible reconstructions and improve consistency of learned representations.

Methods

As a motivating example, we consider a longitudinal dataset of motor function assessments in children with spinal muscular atrophy (SMA). A conditional VAE learns a low-dimensional representation, on which a longitudinal mixed-effects model is fitted to capture disease progression.

To incorporate LLM supervision, reconstructed and original observations are converted into textual descriptions reflecting age-specific SMA motor function profiles and deficiencies. These are presented to an LLM, which performs pairwise comparisons and assigns SMA subtype labels. Multiple pairwise decisions are checked for consistency and aggregated into a final classification.

To enable gradient-based optimisation and minimise LLM calls, a neural proxy model is trained on these LLM outputs to approximate the decision function. Afterwards, we incentivize clinical consistency by minimising the KL-divergence between model input and output via the corresponding LLM-derived SMA class distributions.

Results

Incorporating LLM supervision reduces disagreement between original and reconstructed SMA labels from approximately 12.5% to 6%, depending on the weighting of the supervision term.

We also investigated whether the learned representation becomes more clinically informative by predicting the attainment of external endpoints, specifically functional milestone attainment like sitting and walking. Here, LLM-feedback increased prediction performance when evaluated via Cox regression, outperforming models trained on observed data and learned representations without LLM-guided restructuring.

Discussion

Our results indicate that LLMs can act as synthetic clinical supervisors, steering representation learning toward clinically meaningful structures. Incorporating LLM-derived feedback exposes regions where model misspecification may lead to altered clinical interpretations, enabling targeted correction and improving downstream prediction performance.

LLM-Guided Framework for Privacy-Preserving Discovery of Dynamic Models

14:45-15:00

Presenter: Hanning Yang

Authors: Hanning Yang [1]; Meropi Karakioulaki [2]; Cristina Has [2]; Moritz Hess [1]

Affiliations: [1] Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center - University of Freiburg, Freiburg; [2] Department of Dermatology, Medical Faculty and Medical Center, University of Freiburg, Freiburg im Breisgau

Introduction

Ordinary differential equations (ODEs) are widely used to describe temporal dynamics in biomedical research. However, discovering suitable ODE systems and calibrating their parameters remain challenging, especially for rare diseases with incomplete, sparse, and noisy data like Epidermolysis Bullosa (EB). Existing approaches often rely on direct access to individual-level patient data for model development and parameter estimation, raising concerns about data privacy and availability. Large language models (LLMs) have shown strong reasoning and knowledge integration capabilities. Therefore, we propose LLM-ODE, a framework for ODE discovery and parameter optimization that operates solely on summary statistics of observed data.

Methods

LLM-ODE is an automatic framework consisting of three phases. Phase 1 queries an LLM to propose and refine candidate ODE structures with a structured prompt containing background knowledge. Phase 2 covers parameter inference and optimization. The LLM first infers parameter distributions from domain knowledge and model context, then a semi-autonomous agentic LLM iteratively refines them using available tools to diagnose discrepancies in summary statistics between simulated and empirical data and generate natural language feedback for the next optimization step. Fit is tracked using a log synthetic likelihood (log SL). Phase 3 maintains an experience buffer of ODE structures, parameter distributions, and scores, from which candidates are sampled from top performers and fed back into Phase 1 to guide subsequent structure search. Final top-scoring systems are verified with experts for interpretability and consistency.

Results

Using EB patient data, LLM-ODE progressively aligned synthetic summary statistics with empirical ones across iterations, which was further reflected in trajectory-level fits. In validation, the best-performing ODE structure achieved good trajectory fit across patients, with LLM-inferred parameter distributions showing strong alignment to data-calibrated estimates quantified by KL divergence. All results were obtained without access to individual-level patient records, demonstrating the framework's robustness to the data challenges typical of rare diseases.

Discussion

This work demonstrates the potential of LLM reasoning for knowledge extraction and agentic reflection in a rare disease setting. Future work will focus on extending the framework to capture more complex disease dynamics.

Adaptive and Comparative Trial Design Strategies

Mon, 28.09.26, 13:30–15:00, Room: K5

Chair: Dominic Magirr

Cases against dual primary endpoints: why statistical and clinical criteria should be aligned to guide the design of confirmatory clinical trials

13:30-13:45

Presenter: Nele Henrike Thomas

Authors: Nele Henrike Thomas [1]; Xiaofei Liu [1, 2]; Armin Koch [1]; Anika Großhennig [1]

Affiliations: [1] Institute for Biostatistics, Hannover Medical School, Hannover; [2] Biostatistics and Special Pharmacokinetics Unit, Federal Institute for Drugs and Medical Devices (BfArM), Bonn

Introduction

The clinical development programme of a new medicine intended for marketing authorisation typically includes at least one confirmatory trial to formally demonstrate its efficacy. According to the ICH E9 guideline, one primary endpoint is usually defined for which statistical significance must be shown.1 However, clinical situations exist, where more than one primary outcome is required to evaluate the efficacy of a new drug. For example, if two primary endpoints are defined, there is more than one scenario where the study can be considered a success, requiring the exact conditions for study success being pre-specified. If superiority for both outcomes should be demonstrated for formal proof of efficacy, co-primary endpoints have to be defined and no multiplicity adjustment is necessary.2 However, if superiority for at least one outcome is deemed sufficient, dual primary endpoints are defined, usually with a Bonferroni-split of the type I error (T1E) between endpoints.

Methods

While the co-primary and dual primary endpoint concepts are both statistically correct, i.e. the T1E is controlled, we illustrate that dual primary endpoints are in almost all situations not in line with clinical interpretation of trial outcomes, since a formal study success can be concluded if a significant improvement in one is shown despite a (small) detriment in the other.3 We further identify several clinical scenarios where the dual primary endpoint concept has been proposed but should actually be avoided. Using real-word examples of different indications, we discuss for each scenario why dual primary endpoints may be problematic and what the alternative statistical concept should be that is aligned with clinical decision-making.

Results

The dual primary endpoint concept may ultimately lead to invalid conclusions regarding formal study success and is therefore inconsistent with clinical decision-making. In some cases, it may even result in an unnecessary waste of resources, e.g., through an unnecessary increase in sample size.

Discussion

Therefore, we emphasise that the clinical requirement for formal proof of efficacy should be carefully considered at the planning stage and translated into a statistical framework that appropriately guides the design of confirmatory clinical trials.

References

1. ICH E9 Guideline (1998)

2. EMA Multiplicity Guideline (2002)

3. doi:10.1002/pst.2314

BOP2-DC-FE: Bayesian optimal phase II design with dual-criterion decision making incorporating futility and efficacy-stopping boundaries

13:45-14:00

Presenter: Atsuki Hashimoto

Authors: Atsuki Hashimoto [1]; Belay Birlie Yimer [1]; Kentaro Takeda [2]

Affiliations: [1] Quantitative Science and Evidence Generation, Astellas Pharma Europe Ltd, Addlestone; [2] Quantitative Science and Evidence Generation, Astellas Pharma Global Development Inc, Northbrook, Illinois

Introduction

Phase II oncology trials commonly aim to support efficient go/no-go decisions while maintaining acceptable error rates and patient safety. Hypothesis testing framework with the null hypothesis alone may not be sufficient to establish that a drug is sufficiently clinically effective. Interim monitoring can further improve efficiency by stopping early for futility or efficacy when signals are strong, reducing sample size. Several Bayesian designs have been proposed in Bayesian Optimal Phase II (BOP2) framework, however, existing BOP2 extensions have addressed these needs separately: BOP2-DC [1] enables go/consider/no-go decisions based on both statistical evidence and clinically meaningful benefit, while BOP2-FE [2] adds efficacy stopping in addition to futility stopping to enable earlier declarations of success under strong interim signals.

Methods

We developed BOP2-DC-FE, a Bayesian optimal phase II design that integrates dual-criterion decision making with pre-tabulated futility and efficacy-stopping boundaries by combining the BOP2-DC and BOP2-FE frameworks. The design supports multistage interim analyses with prespecified maximum and interim sample sizes and provides go/consider/no-go decisions based on posterior probabilities reflecting both statistical significance and clinical relevance, while also allowing early stopping for futility and efficacy when interim evidence is sufficiently strong.

Results

Simulation studies showed that BOP2-DC-FE achieved higher (or comparable) Correct Go Rate relative to BOP2-DC while maintaining control of False Go Rate, False No-go Rate, and False Consider Rate. Additionally, incorporating efficacy stopping reduced the average sample size under the alternative hypothesis, reflecting earlier trial termination when strong efficacy was observed at interim looks.

Discussion

By unifying dual-criterion decision making with both futility and efficacy stopping, BOP2-DC-FE strengthens decision quality (statistical evidence plus clinical relevance) while improving efficiency through earlier stopping under clear interim outcomes. This can reduce patient exposure and shorten timelines, especially when treatment effects are strong and rapid development decisions are desired.

[1] Zhao, Y., Li, D., Liu, R., & Yuan, Y. (2021). BOP2-DC: Bayesian optimal phase II designs with dual-criterion decision making. Pharmaceutical Statistics 22 (4):605-618.

[2] Xu, X., Hashimoto, A., Yimer, B. B., & Takeda, K. (2025). BOP2-FE: Bayesian optimal phase II design with futility and efficacy-stopping boundaries. Journal of Biopharmaceutical Statistics, 1-18.

Designing a 3-Arm Phase 3 Oncology Study comparing 2 experimental arms to standard of care: A Simulation Study

14:00-14:15

Presenter: Christina Habermehl

Authors: Christina Habermehl [1]; Heiko Götte [1]; Frank Beier [1]; Anja von Heydebreck [1]; Anja Victor [2]; Kaspar Rufibach [1]

Affiliations: [1] Merck Healthcare KGaA, Darmstadt; [2] Merck KgAa, Darmstadt

Introduction

In a standard 2-arm Phase 3 trial, there is typically one primary comparison. However, in a 3-arm clinical trial there can be 2-3 primary comparisons of interest. This work investigates different testing strategies for comparing two investigational arms against standard-of-care in terms of overall survival (OS) under protection of familywise type I error (FWER).

Methods

The trial design consists of three arms, with sample size calculations based on anticipated OS improvements, targeting 80% power for detecting prespecified alternative hypotheses. We compare three analysis strategies: Bonferroni-Holm, Dunnett, and a global 3-arm test followed by pairwise comparisons. All approaches are special cases of a closed testing procedure and thus control the FWER in the strong sense. They differ, however, in their analysis trigger and local alpha levels. The

Bonferroni-Holm and Dunnett test require adjusted local alpha levels, and events for both comparisons to be reached. The global test uses full alpha for all comparisons, but the 3-arm test serves as a gatekeeper, triggered once events for the first comparison are reached. Using simulations, we investigate the properties of these approaches for timing and power for various alternatives under different scenarios for the treatment effect of each comparison.

Results

Under the planning assumptions, all methods achieve power greater than 80% to reject at least one null hypothesis. Interestingly, neither method is universally superior for all data configurations. Additionally, the global test showed moderate time savings when triggered once events for the first comparison are reached, while Bonferroni-Holm carried a lower risk of missing a clinically relevant effect in an individual comparison when the global test did not reject the null hypothesis.

Discussion

While the global test offers the potential to test all individual hypotheses at full alpha, situations can arise where it is not significant while an individual comparison might be. Therefore, simulation studies should assess this risk when designing such trials. Simulation results should be interpreted in the context of the clinical plausibility of data situations to make an informed choice regarding the appropriate testing strategy. The approach presented here may serve as a framework for designing other multi-arm oncology trials with multiple primary comparisons.

Comparison of prevalent new-user and sequential trials designs for studying the effects of treatment discontinuation

14:15-14:30

Presenter: Elliot McClenaghan

Authors: Elliot McClenaghan [1]; Julie Rouette [2]; Emily Granger [3]; Gwyneth Davies [4]; Ruth Keogh [1]; John Tazare [1]

Affiliations: [1] London School of Hygiene & Tropical Medicine, London; [2] GSK, Montréal; [3] NHS Blood & Transplant, London; [4] Population, Policy & Practice Dept, UCL Great Ormond Street Institute of Child Health, London

Introduction

Estimating causal effects of treatment discontinuation using observational data is of increasing interest and policy relevance. However, alignment of eligibility, treatment assignment and time zero can is challenging. Target trial emulation (TTE) can address these issues but within this framework different study designs and corresponding analysis options are available. These include the prevalent new-user (PNU) and sequential trials designs. Comparisons between these designs are lacking, making it difficult to inform selection and implementation of study designs.

We compare the PNU and sequential trials designs in terms of the causal estimands they target and their implementation, using a case study in cystic fibrosis (CF).

Methods

We emulated a target trial of inhaled corticosteroid (ICS) discontinuation in people with CF using the UK CF Registry, based on the CF-WISE randomised controlled trial (PMID:1655669). For the PNU and sequential trials designs we compared: (i) causal estimand, (ii) time zero definition, (iii) cohort construction and data reuse, and (iv) analysis approach.

Treatment effects were estimated comparing time to first pulmonary exacerbation between ICS discontinuers and continuers and summarised using hazard ratios (HR) and risk differences (RD). Competing events were accounted for throughout.

Results

Among a base cohort of 3842 individuals, 225 discontinuers and 2479 continuers contributed to the PNU analysis. Preliminary findings, under an intention-to-treat approach, suggest ICS discontinuation had no effect on pulmonary exacerbation (HR 1.10, 95% CI: 0.86-1.33), consistent with the CF-WISE trial. The RD at 6 months was 0.03 (95% CI: -0.02-0.09). Results from the sequential trials design will be presented and compared at conference.

Both study designs induce different target populations and causal estimands. Differences in time-zero definition and reuse of individuals across trials lead to distinct analytic cohorts with implications for interpretability, efficiency and benchmarking of effect estimates.

Discussion

PNU and sequential trials designs provide principled approaches to studying treatment discontinuation within the TTE framework. However, design choice determines the estimand, target population, and interpretation of estimates. Explicitly specifying and reporting key design components is recommended to ensure replicable TTEs. This comparison offers practical methodological guidance for selecting, implementing and reporting causal designs for deprescribing questions in pharmacoepidemiology.

Performance of recruitment duration and participant count prediction methods: A Monte Carlo simulation study using proper scoring rules

14:30-14:45

Presenter: Manuela Ott

Authors: Manuela Ott; Malgorzata Roos

Affiliations: University of Zurich, Zurich

Introduction

Timely participant recruitment is critical for the successful completion of medical studies. Overly optimistic specifications of recruitment rates at the design stage remain a well-known challenge, as highlighted by Lasagna's law. Design-stage recruitment analysis (DSRA) involves predicting the recruitment duration and the participant count based on certain assumptions about recruitment rates. DSRA prediction methods include point predictions, distributional predictions such as the classical Poisson model for counts and the Erlang model for recruitment duration as well as Bayesian models. According to Gkioni et al. (2020), only 10% of statisticians used distributional DSRA predictions, preferring simple point predictions. To clarify if this approach is appropriate, it is necessary to compare the predictive performance of different DSRA methods. Our primary performance measure is based on proper scoring rules, a well-established family of measures to assess the performance of predictions.

Methods

A Monte Carlo (MC) simulation study based on the homogeneous Poisson process used the mean interval score (IS) and the (continuous) ranked probability score (RPS) to compare DSRA methods in three scenarios corresponding to large (>30), medium (10 to 30) and small (<10) expected counts. For each scenario, we considered correctly specified and several optimistically specified recruitment rate values.

Results

In all settings considered, Poisson and Erlang distributional predictions performed better than point predictions in terms of mean IS and RPS scores.

Discussion

These results stress the importance of distributional predictions for reliable DSRA and discourage the use of simple point predictions, especially if recruitment rate specifications are overly optimistic. To facilitate application of the DSRA prediction methods considered, we will distribute them as open-source software. Application of this general framework based on proper scoring rules and MC simulations to Bayesian DSRA prediction methods is currently in preparation.

Reference

Gkioni, E., Dodd, S., Rius, R., & Gamble, C. (2020). Statistical models to predict recruitment in clinical trials were rarely used by statisticians in UK and European networks. Journal of Clinical Epidemiology, 124, 58-68. https://doi.org/10.1016/j.jclinepi.2020.03.012

Prediction and Imputation with Missing Data

Mon, 28.09.26, 13:30–15:00, Room: N1

Chair: Roch Giorgi

Prediction with Missing Data: Target Probabilities and Missingness Mechanisms

13:30-13:45

Presenter: Pierre Catoire

Authors: Pierre Catoire [1, 2]; Robin Genuer [1]; Cécile Proust-Lima [1]

Affiliations: [1] University of Bordeaux, Inserm, Bordeaux Population Health Research Center, Bordeaux; [2] Sorbonne Université, Improving Emergency Care (IMPEC) FHU, Paris

Introduction

In statistical inference, consistent estimation with missing data typically requires the 'Missing at Random' (MAR) assumption (1). In prediction however, procedures considered biased for inference, such as unconditional imputation or pattern sub-models (2,3), achieve optimal predictive performance even under 'Missing Not at Random' (MNAR) mechanisms. This apparent contradiction resides in the ambiguity about the prediction target. We propose a formal framework distinguishing two targets: the 'Missingness-Unconditional' (MU) probability conditioning only on the observed predictors, and the 'Missingness-Conditional' (MC) probability conditioning on observed predictors and observation pattern.

Methods

We formalise a classification of missingness mechanisms in prediction and derive conditions for consistent prediction of both targets. Notably, conditions weaker than MAR are sufficient for each: consistent MC prediction is achievable under arbitrary missingness mechanisms, whereas consistent MU prediction is achievable under 'MAR for Predictors - Observed Outcome' (MARX-YO), weaker than MAR. We analyse within this framework common procedures, including Maximum Likelihood Estimation (MLE), Multiple Imputation (MI), inclusion of missingness indicators, pattern submodelling (PS) or inclusion of the missingness as a value of the predictors (MIA). Theoretical results are illustrated through simulations and an application predicting severe trauma in the emergency department.

Results

Procedures ignoring observation pattern (MLE, MI) target MU prediction and are Bayes-consistent under MAR or the weaker MARX-YO assumption. Procedures conditioning explicitly or implicitly on observation patterns (PS, MIA, unconditional imputation, MLE or MI with missingness indicators) target MC prediction and are Bayes-consistent under arbitrary missingness mechanisms including MNAR. In the clinical application, PS performed best, suggesting informative missingness.

Discussion

Missingness in prediction fundamentally differs from inference. Defining the prediction target is required before selecting a method. MC prediction may improve accuracy when missingness is informative, but may be vulnerable to missingness shift if the missingness mechanism diverges between development and deployment. The proposed framework provides structured guidance for method selection in predictive modelling with missing data.

Prediction with Blockwise Missingness via Cross-Modal Substitution and Uncertainty Quantification

13:45-14:00

Presenter: Christoffer Lingjærde

Authors: Christoffer Lingjærde [1]; Sylvia Richardson [2]; Riccardo De Bin [3]; Tero Aittokallio [4, 5]; Basil Ell [6]; Valeria Vitelli [1]

Affiliations: [1] Oslo Center for Biostatistics and Epidemiology, Department of Biostatistics, University of Oslo, Oslo; [2] MRC Biostatistics Unit, University of Cambridge, CB2 0SR; [3] Department of Mathematics, University of Oslo, Oslo; [4] Institute for Cancer Research, Oslo University Hospital, Radiumhospitalet, Oslo; [5] Oslo Centre for Biostatistics and Epidemiology, Domus Medica, University of Oslo, Oslo; [6] Department of Informatics, University of Oslo, Oslo

Introduction

In many biomedical prediction settings, rich multimodal measurements are available during model development. Yet, one or more entire modalities are unavailable at deployment because of cost, logistics, or limited assay availability. This creates a fundamental mismatch: models can exploit multimodal structure during training, but must make predictions from a reduced view of each sample at test time. Existing strategies typically either discard the missing modality altogether or attempt to reconstruct it in full, even when only a small subset of its information is actually needed for accurate prediction.

Methods

We propose Automatic Targeted And Cross-Modal Substitution (ATACMS), a deployment-aware framework for prediction under blockwise missing modality. ATACMS first trains a predictor on the full multimodal training data and identifies the features in the missing modality that are most relevant for the prediction task. At deployment, only this targeted subset is substituted from the observed modality, rather than reconstructing the entire missing modality. Uncertainty is quantified using TabPFN-derived predictive posterior distributions for the substituted features and summarized directly from the spread of these posteriors.

Results

In a controlled simulation study with structured within-modality dependence and nonlinear, non-invertible relationships between modalities, ATACMS improved predictive performance over uni-modal baselines. The resulting uncertainty estimates distinguished settings in which cross-modal substitution was stable from settings in which ambiguity in the missing modality materially limited prediction.

Discussion

ATACMS provides a principled and operationally realistic alternative to full-modality reconstruction in multimodal biomedical prediction. By aligning substitution with the downstream objective, the framework reduces unnecessary reconstruction burden while preserving interpretability. The approach is particularly relevant in omics applications, where high-dimensional assays may be available in development cohorts but unavailable in routine deployment. Ongoing work focuses on evaluation in real multimodal molecular data and comparison with alternative substitution strategies.

Dynamic multivariate imputation of continuous monitoring data

14:00-14:15

Presenter: Eugenia Driusso

Authors: Eugenia Driusso [1]; Maarten van Smeden [1]; Ewout Steyerberg [1, 2]; Nicole Erler [1]

Affiliations: [1] Department of Data Science & Biostatistics, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht; [2] Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, Netherlands, Leiden

Introduction

Continuous monitoring systems in hospital wards generate high-frequency vital sign data to support early detection of patient deterioration. These data are typically summarized using clinical risk scores such as the Modified Early Warning Score (MEWS), which aggregates respiratory rate, heart rate, blood pressure, oxygen saturation, and temperature into a single severity measure. However, such data are frequently incomplete. Currently, missing values are implicitly treated as physiologically normal, leading to underestimation of patient risk and delayed clinical alerts. As accurate reconstruction of missing values over time is essential for reliable use of monitoring data in risk assessment, this study aims to compare different approaches for dynamic imputation of vital signs.

Methods

Common imputation strategies in longitudinal data include carried-forward approaches and patient-specific regression models. Both are straightforward to implement, but they either cannot capture temporal dynamics or require conditioning on the other parameters, which becomes infeasible when multiple vital signs are missing simultaneously. Multivariate imputation by chained equations (MICE)

addresses this by iteratively imputing from a sequence of univariate conditional models and allows the imputation to borrow information across individuals. However, MICE is typically employed in static missing data settings. We adapted this framework to a dynamic imputation setting, in which unobserved measurements were reconstructed sequentially for a new subject as data became available. Population-level parameters were estimated from training data and combined with the individual's history to generate imputations by iteratively updating missing values and random effects. Performance was assessed by comparing imputed and observed vital signs and the derived MEWS score under increasing data sparsity.

Results

Overall, carried-forward approaches demonstrated stable performance over time however, their inability to detect physiological changes limited their utility in alarm systems. Patient-specific regression models, while capturing individual trends, tended to diverge progressively and became inapplicable when multiple vital signs were simultaneously missing. The dynamic imputation approach, by contrast, allowed sequential reconstruction of vital signs under multivariate missingness, providing a flexible framework for handling incomplete trajectories.

Discussion

By combining population-level information with patient-specific histories, the multivariate imputation approach enabled reconstruction of simultaneously missing vital signs. Further investigations will determine the limits within which this approach works reliably.

A new predictive mean matching approach for multiple imputation: evaluation and recommendation

14:15-14:30

Presenter: Jingya Zhao

Authors: Jingya Zhao; Baptiste Leurent; Gareth Ambler

Affiliations: University College London, London

Introduction

Multiple imputation (MI) is a widely used approach for handling missing data. Standard parametric MI methods for continuous variables typically rely on normality assumptions, which are often violated in practice. Predictive mean matching (PMM) is a popular semi-parametric MI method that uses a regression model to obtain predicted values then imputes missing values by sampling from observed values with similar predicted values (with two common matching variations: Type 1 and Type 2). Because PMM imputes using observed values, rather than from a model-based draw, it preserves distributional features of the observed data. However, using PMM can lead to an underestimation of the between-imputation variance, resulting in underestimated standard errors and under-coverage.

Methods

To obtain better estimate of the between-imputation variance we propose a new implementation of PMM with bootstrapping (BS-PMM), with both Type 1 and 2 matchings. For each imputation, a bootstrap sample is drawn from the observed data, and donors are selected from the bootstrap sample using distance-based matching. We compared standard PMM with bootstrap PMM in a simulation study. We simulated a completely observed covariate X and an incomplete outcome Y linearly associated with X. We considered different association between the covariate and the outcome, and different missingness mechanisms: missing completely at random (MCAR), weak missing at random (MAR), and strong MAR. The estimand of interest was the mean of Y and the regression slop of Y on X. We assessed bias, error in model-based standard errors, and confidence interval coverage.

Results

Bootstrap PMM produced larger and more accurate standard errors and generally improved confidence interval coverage compared with standard PMM. However, in some scenarios (weak association between X and Y), BS-PMM could over-estimate the standard errors when used with Type-1 matching. Also, except under MCAR or when there was no association between X and Y, both PMM and BS-PMM tended to return biased point estimates.

Discussion

Bootstrap PMM can improve between-imputation variability relative to standard implementation of PMM, but both approaches remain generally biased under MAR. These findings suggest that PMM MI should be used with caution rather than assumed to perform well by default.

One-step Outcome Imputation: An Alternative to Multiple Imputation

14:30-14:45

Presenter: Andreas Nordland

Authors: Andreas Nordland; Klaus Kähler Holst; Aske Thorn Iversen

Affiliations: Novo Nordisk A/S, Søborg

Introduction

Missing outcomes in randomized controlled trials (RCTs) are often handled by multiple imputation (MI). Rubin's rule is routinely used to estimate standard errors, but can fail to provide valid variance estimates for some commonly used procedures, such as reference-based imputation (Bartlett, 2023). We propose an alternative: explicitly formulating the estimand/target parameter implied by a given imputation model and constructing an efficient one-step estimator and its influence function. This approach yields valid asymptotic inference even when Rubin's rule is inconsistent, and circumvents the stochastic nature and computational complexity of MI. The method is implemented in the targeted R package for commonly used imputation procedures.

Methods

We generalize and exemplify the work of Robins & Wang (2000) in an RCT context and focus on estimating the imputation average treatment effect (ATE). We make a natural decomposition of the target parameter into the non-missing outcomes ATE and the missing outcome imputation ATE. This decomposition allows us to use a nonparametric (efficient) estimator for the non-missing outcome ATE. The imputation model can be any function of baseline covariates, the treatment, and post-randomization variables. The only requirement is that the imputation model has a known parametrization and that the estimator for the parameter is regular and asymptotically linear. The estimators and associated influence functions are finally combined using the delta method for influence functions. Interestingly, if the imputation function depends on post-randomization variables, efficient estimation will require a working model for the imputation function marginalized over the post-randomization variables.

Results

We provide one-step estimators for a range of imputation procedures used in RCTs and compare the performance (statistically and computationally) to existing multiple imputation procedures in simulation studies.

Discussion

Robins, James M., and Naisyin Wang. 'Inference for imputation estimators.' Biometrika 87.1 (2000): 113-124.

Bartlett, Jonathan W. 'Reference-based multiple imputation - what is the right variance and how to estimate it.' Statistics in Biopharmaceutical Research 15.1 (2023): 178-186.

A Causal Inference Framework for Antihypertensive Treatment Selection Using Pathway-Aware Genetic Profiles in UK Biobank

14:45-15:00

Presenter: Jaegyun Jung

Authors: Jaegyun Jung [1]; Hyung-Chul Lee [1, 2]

Affiliations: [1] Seoul National University Hospital, Seoul; [2] Seoul National University, Seoul

Introduction

In essential hypertension, angiotensin-converting enzyme inhibitors (ACEi) and calcium channel blockers (CCB) are both acceptable first-line monotherapy options, yet routine prescribing provides limited guidance on which class is more likely to benefit an individual patient. We therefore developed a biologically informed, pathway-aware framework that constructs patient-level genetic profiles and applies causal inference to estimate whether ACEi or CCB is expected to provide greater cardiovascular benefit for each individual.

Methods

Using linked UK Biobank genotype data, electronic health records, and primary care prescribing data, we analyzed 44,713 participants initiating ACEi or CCB. We grouped 44 blood-pressure-associated SNPs into four biologically defined mechanism classes: renin-angiotensin-aldosterone system (RAAS), calcium channel, vascular system, and renal/natriuretic system, and derived continuous pathway-specific genetic burden profiles from allele-aligned dosages. We then applied targeted maximum likelihood estimation with Super Learner to estimate patient-level treatment benefit for 5-year major adverse cardiovascular events (MACE), compare genetically preferred versus actually prescribed therapy, and quantify counterfactual risk reduction under profile-concordant treatment.

Results

Actual ACEi versus CCB prescribing was essentially unrelated to genetically inferred treatment preference, with near-even cross-classification between prescribed and genetically preferred therapy. Despite this, patients who received the genetically preferred drug had 16% lower observed odds of 5-year MACE than those who received the non-preferred drug. Counterfactual analyses showed that patients who received a drug class discordant with their pathway-aware genetic profile would have had approximately 13% lower 5-year MACE risk if treated with the profile-concordant drug, with bootstrap 95% confidence intervals ranging from 6.3% to 24.7% across the two switch directions. The dominant biological driver of treatment differentiation was the RAAS profile.

Discussion

A biologically informed, pathway-aware genetic framework identified clinically meaningful patient-level differences in optimal antihypertensive therapy that were not captured by routine prescribing. These findings support genotype-guided antihypertensive selection and identify RAAS as the principal axis underlying differential benefit between ACEi and CCB.

Health Data Infrastructure and Secondary Data Use

Mon, 28.09.26, 13:30–15:00, Room: N2

Chair: Andrea Mario Vergani

The European Health Data Space and the Secondary Use of Sensitive Health Data

13:30-13:45

Presenter/Author: Sameeksha Sriram, Sin-Yin Huei, Marie Haucke, Ulrich Sax

Affiliations: University of Göttingen, Campus-Institute Data Science (CIDAS), Goettingen; University of Göttingen, Göttingen; Institut für Medizinische Informatik, Universitätsmedizin GöttingenCampus-Institut für Data Science (CIDAS), Georg-August-Universität Göttingen, Göttingen

Introduction

The European Health Data Space (EHDS) is a major EU data-governance initiative designed to enable cross-border access to and reuse of electronic health data for care, research, innovation, and public-interest purposes. It aims to reduce fragmentation, improve interoperability, and make health data more reusable across Member States. However, its success depends not only on technical infrastructure but also on legal, ethical, and organizational harmonisation, especially for genetic and genomic data. This paper examines the EHDS with a focus on secondary use, outlines its institutional architecture, discusses Finland's Findata as a national model, and analyzes challenges for data holders and data donors, including interoperability, governance burden, privacy, and re-identification risk. Through a cross-border cancer genomics case study, it argues that while the EHDS can improve data discovery and access procedures, it does not by itself resolve the legal and ethical fragmentation that still hinders cross-border genomic research.

Methods

All authors participated in the Seminar Series Good Practice Methods of Privacy and Ethics in Data Science - on the sub topic of "the impact of data sharing - EHDS: [what are the chances for research, which could be the problems for data holders/data donors]. This included primarily a literature search on previously published data use example, the infrastructure and regulatory affairs. Secundarily a seminar presentation and discussion including a ten page manuscript which was used in this paper.

Results

The paper next analyzes challenges from the perspectives of data holders and data donors, focusing on interoperability, governance burdens, privacy protec- tion, residual re-identification risk, and the special sensitivity of genomic data. Finally, it develops the argument through a cross-border cancer genomics case study, showing that although the EHDS can streamline discovery and routing of access requests, it does not by itself eliminate the legal fragmentation, ethics heterogeneity, and consent-related barriers that currently impede cross-border genomic research.

Discussion

EHDS improves structured, secure health-data reuse, but genomic research still faces fragmented consent, ethics, interoperability, and linkage rules infrastructure alone will not deliver true cross-border harmonisation.

The full paper of this abstract will be published in the GMDS series "Studies in Health Technology and Informatics" with doi: 10.3233/SHTI261011, URL:https://dx.doi.org/10.3233/SHTI261011

Enabling the flexible integration of trusted third party tools in OSSE registries - a technical case report

13:45-14:00

Presenter: Jessica Vasseur

Authors: Jessica Vasseur [1]; Jens Göbel [1]; Axel Zieschank [1]; Sabine Hanß [2]; Anne Pelz [3]; Holger Storf [1]

Affiliations: [1] Goethe University Frankfurt, University Medicine, Institute of Medical Informatics (IMI), Frankfurt am Main; [2] University Medicine Göttingen, Institute for Medical Informatics, Göttingen; [3] Institute for Medical Informatics and Biometry, Faculty of Medicine Carl Gustav Carus, TUD Dresden University of Technology, Dresden

Introduction

Medical registries and their electronic data capture systems, such as the Open Source Registry System for Rare Diseases (OSSE), require integration of trusted third party (TTP) services to meet data protection requirements. While OSSE is currently coupled to the Mainzelliste, deployment in research infrastructures such as the Network University Medicine (NUM) demands more flexible support for heterogeneous TTP tools.

Methods

We developed OSSE TTP, a configurable adapter mediating communication between the OSSE EDC and TTP services. The modular design maintains compatibility with existing OSSE-based registries and workflows while enabling integration into established infrastructures such as NUM NUKLEUS.

Results

The OSSE TTP prototype supports selection of the Mainzelliste or E-PIX/gPAS for identity management and pseudonymization, with optional connection to gICS for consent management. It provides the necessary endpoints and user interfaces for seamless integration into existing workflows and authorization concepts. A demonstrator was successfully developed and will be validated prior to gradual deployment within the NUM4Rare infrastructure.

Discussion

OSSE TTP extends the applicability of OSSE through flexible, service-agnostic TTP integration, building on existing processes and interfaces to ensure data protection-compliant data exchange. The modular design can accommodate future integration of additional TTP services. Overall, OSSE TTP provides a flexible and data protection-compliant foundation for connecting various TTP services to OSSE, representing a key step toward the integration of OSSE-based registries into research infrastructures.

The full paper of this abstract will be published in the GMDS series "Studies in Health Technology and Informatics" with doi: 10.3233/SHTI260981, URL:https://dx.doi.org/10.3233/SHTI260981

Supporting Complex Adaptive Trial Designs in an EDC System: Multiple Randomizations, Biomarker Stratification, and Platform Trial Structures

14:00-14:15

Presenter: Anita Popp

Authors: Manuel Neukum [1]; Anita Popp [1, 2]

Affiliations: [1] AB Cube Germany GmbH, Munich; [2] AB Cube Germany, Munich

Introduction

The advent of precision medicine trials is characterized by a shift from traditional tumor-centered to gene-directed and histology-agnostic approaches. This shift increasingly requires EDC systems to handle complex randomization logic beyond simple parallel-group allocation in former RCTs (Randomized-Clinical Trials). Basket, umbrella, and platform trials demand dynamic, multi-layered randomization schemes, including biomarker-driven stratification, response-adaptive allocation, and re-randomization of individual patients across treatment phases. The technical feasibility of implementing such designs within a validated, regulatory-compliant EDC system remains a practical challenge for sponsors and data managers.

Methods

We herein describe the implementation tool-set offor complex adaptive randomization workflows available within Marvin EDC. The system's randomization engine supports stratified randomization, permuted block randomization, covariate-adjusted response-adaptive randomization (CARA), and multiple sequential randomizations per patient within a single trial. Configuration is performed through Marvin's metadata-driven design environment without custom programming, preserving validation status across protocol amendments. A real-world interventional trial with a multi-arm adaptive structure is presented as an anonymized use case, illustrating how treatment arms and stratification variables were configured, modified during the study, and maintained within a complete audit trail.

Results

The described trial was successfully conducted using Marvin EDC's adaptive randomization capabilities. Multiple per-patient randomization events were managed within a single study instance, with each allocation documented in a tamper-proof audit trail. Biomarker-based stratification variables were integrated directly into the randomization algorithm. Arm-level modifications during the trial were implemented without revalidation of the core system, and CDISC-compliant data exports were maintained throughout the entire process.

Discussion

Marvin EDC demonstrates that adaptive and master protocol trial designs, including platform trials with dynamic arm management and multiple per-patient randomizations, are operationally viable within a commercially validated and EMA/ FDA standards-compliant EDC platform. The metadata-driven configuration approach reduces implementation effort and supports regulatory submissions without compromising system integrity. This positions Marvin EDC as a practical solution for sponsors pursuing precision medicine trial strategies

---

Bridging the Gap: Leveraging dbplyr for Efficient, Portable, and Scalable Analysis of German SHI Claims Data in the FDZ Gesundheit Secure Processing Environment

14:15-14:30

Presenter: Jonas Krampe

Authors: Jonas Krampe; Sima Melnik

Affiliations: Gesundheitsforen Leipzig GmbH, Leipzig

Introduction

With the launch of the Forschungsdatenzentrum (FDZ) Gesundheit, researchers now have access to nationwide German Statutory Health Insurance (SHI) claims data encompassing millions of insured individuals. The data is hosted on a high-performance SAP HANA database within a Secure Processing Environment (SPE). However, the complex data models (e.g., DM1-2 to DM3 transitions) and the necessity to write highly optimized SQL to handle tables with billions of rows (such as AMBDIAG or REZ) present a steep technical hurdle for epidemiologists and biostatisticians accustomed to R. To address this, our objective is to present a scalable and reproducible workflow using the R package dbplyr as a translation layer between familiar tidyverse syntax and HANA-optimized SQL, effectively lowering the technical barrier for health data scientists.

Methods

We developed a standardized analysis pipeline tailored to the FDZ system architecture. Using dbplyr, we establish a 'Lazy Evaluation' workflow where complex data manipulations - including cohort definitions, comorbidity indexing, and longitudinal follow-ups - are translated into SQL and executed

entirely server-side. Furthermore, we demonstrate that scripts can be initially developed and tested deploying the Public Use File locally on a local SAP HANA database. Additionally, if other claims data exist on different SQL backends, they can also be accessed via the dbplyr method. Utilizing the exact same code to assess both the FDZ data and external databases provides a powerful method to verify the script not only technically, but also to confirm it is analytically correct.

Results

The dbplyr workflow prevents memory exhaustion in the FDZ virtual analysis rooms by ensuring only aggregated or strictly filtered result sets are collected into the local R session. This approach significantly reduces query execution times and data transfer bottlenecks. The abstraction also simplifies handling schema changes between DM1-2, DM3, and DM4.

Discussion

Utilizing dbplyr at the FDZ Gesundheit empowers researchers to conduct large-scale claims data analyses without requiring advanced SQL expertise. By decoupling the analysis logic from the database backend, this workflow fosters highly reproducible research, facilitates smooth transitions from development to production, and aligns with the modern demands of processing massive biomedical datasets. Furthermore, it provides a transparent mechanism for technical validation and compliance with FDZ 'Code-as-Documentation' requirements.

Development of a generic method to describe care pathways of patients with rare diseases using the French National Health Data System (SNDS).

14:30-14:45

Presenter: Renée Le Clech

Authors: Renée Le Clech [1]; Aurélie Bannay [2]; Nicolas Garcelon [3]; Nicolas Jay [2]; Anne-Sophie Jannot [4, 5]

Affiliations: [1] Université Paris Cité, Inria, Inserm, HeKA, F-75015, Paris; [2] Medical Assessment and Information Department, Université de Lorraine, CHRU-Nancy, Nancy; [3] Imagine - Institut des Maladies Génétiques, Paris; [4] HeKa Université Paris Cité, INRIA, INSERM, F-75015, Paris; [5] Banque Nationale de Données Maladies Rares (BNDMR), AP-HP, Paris

Introduction

Processing medico-administrative databases for research purposes presents challenges given the diversity of data sources and the complexity of the information they contain. The French National Health Data System (SNDS) includes over 38,000 medical codes describing patients medical pathways, including diagnoses, medications, biological acts and medical devices. For rare diseases, patient pathways are heterogeneous, and an understanding of real-world global care management represents a major step forward for research. We use the SNDS extracted on a rare disease population to develop a method that describes patient typical pathways using a case-control comparison.

Methods

We propose a method to estimate typical care pathways for patients with rare diseases by summarizing them into a set of key "medical care events." First, we select a population of patients with data available over a comparable time period. Second, we identify discriminative medical codes across multiple classifications, including hospital diagnoses (ICD-10), medications (ATC), medical procedures and biological tests recorded in specific medico-administrative classifications. Third, we apply clustering methods, k-means and agglomerative hierarchical clustering (AHC), using a precomputed embedding matrix estimated on a 3 million random sample from the SNDS. Clustering robustness is assessed with a bootstrap analysis, and stability is quantified using the Adjusted Rand Index (ARI). The method is applied on a cohort of patients with Dravet syndrome (n_case = 125, n_control = 326).

Results

Among the 2189 medical codes, 207 codes were identified as discriminative. AHC, which showed the highest stability (ARI = 0.88 ± 0.06), was performed on these 207 codes using the embedding matrix, and resulted in 21 clusters, corresponding to 21 "medical care events." The clusters group SNDS codes reflecting similar health events while coming from different classifications, highlighting the value of embedding-based approaches. For example, a cluster grouped epilepsy CIM-10 codes G40 and G41, ATC code N03A (antiepileptic drugs), and relevant medical procedure codes, despite belonging to different classifications.

Discussion

This method condenses hundreds of medical codes into a limited number of meaningful "medical care events," simplifying the analysis of patient pathways. It provides a generic, unsupervised framework for exploring real-world care management in rare diseases.

Early Postoperative AKI Risk Factors After Pediatric Cardiac Surgery: A Literature-Informed Analysis of Integrated Routine Clinical Data

14:45-15:00

Presenter: Fatemeh Ali-Sadeh

Authors: Johanna Apfel-Starke [1, 2]; Marcel Schack [3]

Affiliations: [1] Medizinische Hochschule Hannover Hochschule Hannover, Hannover; [2] Hannover Medical School, MHH Information Technology, Hannover, Germany, Hannover; [3] Peter L. Reichertz Institute for Medical Informatics, TU Braunschweig and Hannover Medical School, Hannover, Germany, Hannover

Introduction: Acute kidney injury (AKI) is a frequent complication after pediatric cardiac surgery. This study aimed to identify literature-informed perioperative variables associated with AKI within 0-7 days after surgery and to assess whether the resulting risk profile remained visible in an extended routine-data feature set. Methods: We performed a retrospective secondary analysis of anonymized

routine clinical data comprising 1,354 surgery episodes from 1,135 patients. Feature selection was informed by a structured literature review and resulted in two analytical datasets from the same cohort: Dataset A, a 26-feature core set, and Dataset B, a broader 72-variable feature space used for domain-level robustness analysis. The primary endpoint was AKI within 0-7 days after surgery. Results: AKI within 0-7 days occurred in 530 of 1,354 episodes (39.14%), and overall postoperative AKI occurred in 566 episodes (41.8%). Episodes with AKI showed longer procedure, cardiopulmonary bypass (CPB), and aortic cross-clamp times and higher mean Vasoactive Inotropic Score (VIS) during the first 24 postoperative hours. The median time from surgery end to the first linked rule-based AKI event was 7.2 hours 395 of 530 events (74.5%) occurred within 24 hours and 462 (87.2%) within 48 hours. Discussion: The findings support an early postoperative risk profile involving patient vulnerability, intraoperative burden, and early hemodynamic/perfusion instability. Dataset B provided domain-level robustness support. Conclusion:Literature-informed feature selection combined with integrated routine clinical data can reveal clinically interpretable perioperative AKI risk domains after pediatric cardiac surgery. ehrapy and AnnData enabled reproducible data representation and exploration.

The full paper of this abstract will be published in the GMDS series "Studies in Health Technology and Informatics" with doi: 10.3233/SHTI260992, URL:https://dx.doi.org/10.3233/SHTI260992

Diagnostic Evidence Synthesis and Test Evaluation

Mon, 28.09.26, 13:30–15:00, Room: N4

Chair: Bethany Hamilton Bhat

A framework for synthesizing evidence from human and animal studies: the GALENOS approach to triangulating evidence

13:30-13:45

Presenter: Virginia Chiocchia

Authors: Virginia Chiocchia [1]; Andrea Cipriani [2, 3]; Georgia Salanti [4]

Affiliations: [1] University of Bern, Bern; [2] Oxford Precision Psychiatry Lab, Department of Psychiatry, University of Oxford, Oxford; [3] Oxford Health NHS Foundation Trust, Oxford; [4] Institute of Social and Preventive Medicine, University of Bern, Bern

Introduction

The Global Alliance for Living Evidence on aNxiety, depressiOn and pSychosis (GALENOS), launched in 2023, is an international interdisciplinary initiative designed to advance open, transparent, and reproducible research synthesis in mental health. GALENOS aims to develop and maintain an open-access, living evidence resource to inform early-phase research prioritisation for anxiety, depression, and psychosis. With the contribution of researchers, methodologists, clinicians, and people with lived experience, this project integrates and synthesises findings from different sources of evidence, e.g. preclinical and clinical studies, on both pharmacological and non-pharmacological treatments, as well as diagnostic and prognostic tools.

Methods

Central to GALENOS is the adoption and development of innovative, fit-for-purpose methodologies for living systematic reviews (LSRs) and meta-analyses. Particularly, we developed a framework to synthesise evidence from different sources, specifically for human and animal studies. The framework covers living review workflows, meta-analytic methods tailored for the specific data, and procedure for summarising and evaluating the biases and credibility of the results from these different sources. The main output of interest is documented in detailed summary of evidence tables, which include the summary of the association, internal and external validity issues, as well as reporting bias, and other considerations.

Where applicable, triangulation methods are used to integrate the findings from the distinct approaches, accounting for both the direction of the biases and the strength of evidence. An independent multidisciplinary panel determines whether there are consolidating conclusions or shares further insights to inform the next update of the review.

Results

At the time of abstract submission, GALENOS has produced nine LSRs (either published or in progress) and has held three triangulation meetings.

We present the methodologies involved in the process and the challenges identified in their implementation. In particular, we describe how our proposed framework fits into the structured, multidisciplinary GALENOS approach to triangulating evidence (GATE) that was developed as part of the programme.

Discussion

Integrating different types of evidence within a living, co-produced framework can strengthen early-phase research decisions and accelerates translation from discovery science into effective new treatments and systems. By creating open-access research outputs, the GALENOS approach supports transparent, evidence-based research prioritisation.

A new generation of statistical models exploits the full information of multiple thresholds in the meta-analysis of diagnostic test accuracy studies

13:45-14:00

Presenter: Ferdinand Valentin Stoye

Authors: Alexander Hapfelmeier [1, 2]; Ferdinand Valentin Stoye [3]; Annika Hoyer [3]; Antonia Zapf [4]; Alex J Sutton [5]; Klaus Linde [1]; Antonius Schneider [1]

Affiliations: [1] Institute of General Practice and Health Services Research, TUM School of Medicine and Health, Technical University of Munich, Munich; [2] Institute for AI and Informatics in Medicine, TUM School of Medicine and Health, Technical University of Munich, Munich; [3] Biostatistics and Medical Biometry, Medical School OWL, Bielefeld University, Bielefeld; [4] Institute of Medical Biometry and Epidemiology, University Medical Center Hamburg-Eppendorf, Hamburg; [5] Biostatistics Research Group, Division of Public Health & Epidemiology, School of Medical Sciences, University of Leicester, Leicester

Introduction

Meta-analyses of diagnostic test accuracy studies typically use a standard bivariate random-effects model to estimate sensitivity and specificity using a single diagnostic threshold for an ordinal or metric test result per primary study. Although this aligns with present recommendations, a new generation of statistical models has emerged that incorporates information from all thresholds.

Methods

We briefly introduce four exemplars of the novel models (linear mixed model by Steinhauser et al. [1], time-to-event model by Hoyer et al. [2], Bayesian mixed model by Jones et al. [3], semiparametric model by Frömke et al. [4]). We illustrate their properties and benefits on a use case on screening for anxiety disorders [5] and discuss implications for clinical practice.

Results

By contacting primary study authors, published results from 12 studies on screening for generalized anxiety disorder using the questionnaire GAD-7 were completed with results on additional thresholds. Threshold-dependent estimated summary sensitivities and specificities are comparable for the multiple threshold models and mostly consistent with published results (mean sensitivity (specificity) difference between 0.009 and 0.071 (0.007 and 0.023)). In contrast, the standard model, estimated separately on each threshold, deviates more between both variants and violates monotonicity assumptions in sensitivity and specificity (mean difference in sensitivity (specificity) 0.088 (0.044)). This also translates into estimates of positive/negative predictive values.

Discussion

Simulation studies and empirical evidence have shown that novel models mitigate reporting bias and estimate diagnostic test accuracy with high precision, even for incomplete threshold information. Exploring diagnostic test accuracy across multiple thresholds also increases the clinical usefulness of such meta-analyses, as different thresholds may be relevant in different situations, e.g., to rule in or out a target condition.

[1] Steinhauser S, Schumacher M and Rücker G. BMC Med Res Methodol 201616:97. DOI: 10.1186/s12874-016-0196-1

[2] Hoyer A, Hirt S and Kuss O. Res Synth Methods 20189:62-72. DOI: 10.1002/jrsm.1273

[3] Jones HE, Gatsonsis CA, Trikalinos TA, et al. Stat Med 201938:4789-4803. DOI: 10.1002/sim.8301

[4] Frömke C, Kirstein M and Zapf A. Res Synth Methods 2022 13: 612-621. DOI: 10.1002/jrsm.1579

[5] Aktürk Z, Hapfelmeier A, Fomenko A, et al. Cochrane Database Syst Rev 2025 3:CD015455. DOI: 10.1002/14651858.CD015455

Meta-analysis for the comparison of multiple diagnostic tests to a common reference standard

14:00-14:15

Presenter: Jannik Rehse

Authors: Jannik Rehse; Ferdinand Valentin Stoye; Annika Hoyer

Affiliations: Biostatistics and Medical Biometry, Medical School OWL, Bielefeld University, Bielefeld

Introduction

Meta-analysis of diagnostic test accuracy studies is an important area of biostatistical research, with applications in the comparison of multiple diagnostic tests against a common reference standard and the aim of comparing test performance in terms of sensitivity and specificity. Standard approaches use a common study-specific effect across all tests, but do not fully account for correlations between tests within primary studies, which may lead to biased estimates. Hoyer et al. [1] proposed a

model that explicitly accounts for these correlations between two tests. We generalize this framework to multiple tests, allow studies to include only subsets of these tests, and consider multiple diagnostic thresholds for each test.

Methods

We extend the quadrivariate generalized linear mixed model (GLMM) of Hoyer et al. for the joint analysis of multiple diagnostic tests. The model is formulated as a multivariate GLMM including test-specific fixed effects and a multivariate random effects structure to capture between-study heterogeneity and within-study correlations of the tests. The resulting covariance structure enables appropriate quantification of uncertainty and valid inference for differences between tests in terms of sensitivity, specificity and AUC. Threshold values are included as covariates in the linear predictor to account for their effect on test accuracy and to enable estimation and comparison of ROC curves.

Results

Simulation studies suggest that the proposed model provides unbiased estimates of sensitivity and specificity and enables valid comparisons between tests, with appropriate empirical coverage. Due to the increased number of parameters, convergence issues may arise in settings with few studies or small sample sizes. The method will be applied to a real-world dataset on screening for anxiety disorders, including approximately 50 studies in which up to six diagnostic tests were evaluated.

Discussion

The proposed framework extends existing methods to multiple tests and incomplete data. Limitations of applicability and robustness are investigated in a simulation study. Overall, the model enables coherent comparisons by appropriately accounting for dependencies between tests.

[1] Hoyer A, Kuss O. Meta-analysis for the comparison of two diagnostic tests to a common gold standard: A generalized linear mixed model approach. Stat Methods Med Res. 201827(5):1410-1421. DOI: 10.1177/0962280216661587

Subgroup Analysis in Diagnostic Test Accuracy Meta-Analysis with Mixed-Model Trees (DamasTree)

14:15-14:30

Presenter: Lukas Mischinger

Authors: Lukas Mischinger; Alexander Hapfelmeier

Affiliations: TUM University Hospital, Munich

Introduction

Investigating sources of heterogeneity is an important step in diagnostic test accuracy (DTA) meta-analysis. However, existing guidance and software mainly focus on meta-regression comparing pre-specified subgroups defined by a single binary covariate. To address these limitations, we propose

multivariable modelling by damastree, a method that combines mixed-effects modeling with recursive partitioning to automatically identify heterogeneous subgroups in the data. The approach accommodates multi-level categorical and continuous covariates and can detect interactions.

Methods

We developed damastree for investigating heterogeneity in DTA meta-analysis. The method estimates summary sensitivity and specificity using a binomial mixed model with study-specific random intercepts and slopes, analogous to the bivariate random-effects model. It adapts generalized linear mixed-model trees to identify subgroups with heterogeneous fixed effects. An accompanying R package returns node-specific accuracy estimates, visualizations in summary ROC space and forest plots, and derived measures of heterogeneity and diagnostic performance.

Results

We evaluated damastree in simulations assessing type-II and type-I error and in a real data application. In type-II simulations, data were generated to contain three subgroups defined by an interaction between two of five covariates of different scale. Preliminary results indicate that recovery of the true subgroups improves with increasing numbers of primary studies and stronger heterogeneity, while it worsens with greater unexplained between-study variability and threshold variability. Under a null scenario without true subgroups, false-positive splitting occurred infrequently and was often below the nominal 5% level, suggesting slightly conservative behavior. In an application to a DTA meta-analysis of the GAD-7 screening instrument for generalized anxiety disorder, damastree identified clinical setting as a relevant source of heterogeneity, yielding three subgroups (mental health settings, specific disease populations, and non-healthcare or unselected populations) with differing sensitivity-specificity trade-offs.

Discussion

These preliminary results suggest that damastree can detect meaningful subgroup structures while maintaining a low type-I error probability. The method provides a flexible framework for exploring complex multivariable sources of heterogeneity in DTA meta-analysis beyond traditional prespecified meta-regression approaches.

Simultaneous Tests and Confidence Intervals for Contrasts of Multiple Maximum Youden Indices in Factorial Diagnostic Trials

14:30-14:45

Presenter: Dirk Schomburg

Authors: Dirk Schomburg; Frank Konietschke

Affiliations: Charité - Universitätsmedizin Berlin, Berlin

Introduction

Diagnostic tests are important for clinical decision-making. A well-known method for a single biomarker is the use of the maximized Youden index and its corresponding cutoff value. However, many potential biomarkers and variations of other involved diagnostic factors can be considered, and their evaluation in a diagnostic trial with a factorial design is required to identify the optimal method for answering the clinical questions. We propose the use of the max-T test as a multiple contrast test, based on ideas from Bretz et al. (2001), to control the family-wise error rate in the strong sense and to exploit the correlations between the test statistics.

Methods

We implemented and evaluated different methods to estimate the multiple maximum Youden indices, their contrasts, and their variance-covariance matrix, which are required in subsequent calculations for local and global null hypothesis tests and simultaneous confidence intervals. Our parametric models include (a) normal distribution, (b) power-normal distribution with a common Box-Cox transformation, and (c) a model allowing for different skewness in diseased and non-diseased groups. We employ the trust-region method (Sorensen, 1982) to numerically maximize the likelihood function. The Hessian matrix is also used to estimate the variance-covariance matrix of the estimates. Some calculations use extended-precision arithmetic and small-sample-size optimizations. Nonparametric approaches include (d) naive estimation and (e) kernel-based estimation methods.

Results

Fast and reliable optimization was achieved with the trust-region method using analytically derived gradients and Hessian matrices, extended-precision arithmetic, code parallelization and further optimization.

Discussion

We discuss the properties of the methods, such as their control of the type I error rate and the power obtained from extensive simulation studies.

References

Bantis LE, Nakas CT, Reiser B. (2021) Statistical inference for the difference between two maximized Youden indices obtained from correlated biomarkers. Biometrical Journal 63, 1241-1253.

Bretz F, Genz A, Hothorn LA. (2001) On the numerical availability of multiple comparison procedures. Biometrical Journal 43(5), 645-656.

Sorensen, D. C. (1982). Newton's Method with a Model Trust Region Modification. SIAM Journal on Numerical Analysis 19, 409-426

Reconstructing individual patient data from published ROC and DCA plots

14:45-15:00

Presenter: Junfeng Wang

Authors: Junfeng Wang [1]; Kim Zhipei Wang [1]; Laure Wynants [1, 2]

Affiliations: [1] Maastricht University, Maastricht; [2] KU Leuven, Leuven

Introduction

Clinical prediction models estimate individual risk of diagnostic or prognostic outcomes. Performance evaluation requires predicted probabilities and observed outcomes to assess discrimination, calibration, and clinical utility. However, most studies do not provide this, and instead report receiver operating characteristic (ROC) curves and decision curve analysis (DCA) plots. We aimed to develop a method to reconstruct individual risk predictions using only published ROC and DCA plots.

Methods

We propose a framework based solely on ROC and DCA figures. Sensitivity-specificity pairs are digitized from ROC curves, and threshold probabilities and net benefit values are extracted from DCA plots. Using the mathematical relationships linking sensitivity, specificity, prevalence, threshold probability, and net benefit, we recover the thresholds underlying ROC coordinates and generate a dataset from which individual predicted probabilities and binary outcomes consistent with the published curves are reconstructed.

The method is implemented in R, and simulations were performed to assess the performance of different function options (e.g. forward/backward/two-pass, strict monotonicity) under various data generation mechanisms (varying AUC, shape of ROC, sample size, event rate, smoothing of decision curve). We also illustrate the method using a published evaluation of the ADNEX model.

Results

The simulation results showed that it is better to not require strict monotonicity, especially when sample size is relatively large (>1000). The two-pass approach had the best performance in almost all scenarios (mean RMSE = 0.0653 in non-strict monotonicity, range 0.03979 to 0.08349) than forward or backward. The ADNEX data could be reconstructed well (RMSE = 0.0143), enabling derivation of calibration plots and additional performance measures that matched those derived from the true data.

Discussion

Reconstructing individual predictions from ROC and DCA plots addresses a key limitation in current reporting practice. The proposed method and R function enables to reconstruct patient data and unreported performance measures and plots. This is useful for critical appraisal and meta-analysis of published studies.

Survival Prediction Models

Mon, 28.09.26, 13:30–15:00, Room: Runder Saal

Chair: Rebecca Whittle

Regression vs. Machine Learning for Variable Selection in Clinical Prediction Modeling: A Simulation Study

13:30-13:45

Presenter: Johannes Vey

Authors: Johannes Vey [1]; Georg Heinze [2]; Meinhard Kieser [1]

Affiliations: [1] Institute of Medical Biometry, University of Heidelberg, Heidelberg; [2] Medical University of Vienna, Vienna

Introduction

A wide range of methods exist for developing clinical prediction models (CPM). In addition to regression models, machine learning algorithms have emerged as powerful prediction tools over the last three decades. For predicting clinical outcomes, it is important not only to decide which model is applied but also to select the most relevant variables. Variable selection can improve model performance, interpretability, and generalizability.

Methods

In this simulation study, we conducted a neutral comparison of different regression-based (linear regression with stepwise selection, regularized regression) and machine learning (gradient boosting, random forest) variable selection strategies to derive a CPM predicting a continuous outcome. The generated datasets included a total of 15 variables, with 8 of those being predictor variables. Four data-generating mechanisms with increasing complexity, of which one is based on the design proposed by Binder et al. 2011, produced data structures typical for biomedicine covering linear associations and gradually introducing non-linear and non-additive elements.

Results

All methods generally performed better with increasing sample size and less noise in the data. Gradient boosting with regression models (GBM) and with trees (GBT) as base learners, and the elastic net regularized regression (ENET) selected nearly all variables, especially with increasing sample size. The linear regression model with stepwise selection (LMSS) balanced the inclusion of predictors and exclusion of non-predictor variables well. In more complex data, variable selection using the Boruta (RFB) and Hapfelmeier (RFH) approaches for random forests performed similarly as LMSS. There were no considerable differences in the predictive performance in test data between the methods, whereby the regression-based LMSS, ENET, and GBM had slightly more accurate predictions in smaller datasets of up to 500 observations.

Discussion

A sufficiently large dataset should be used to develop a CPM and to identify the important predictors, with LMSS being better at handling smaller datasets. With larger (from 500-1000 observations) and more complex data, random forests with the Boruta or Hapfelmeier approaches for variable selection are reasonable tree-based alternatives to LMSS.

References:

Binder, H., Sauerbrei, W., Royston, P. Multivariable Model-Building with Continuous Covariates: 1. Performance Measures and Simulation Design. Freiburg: University of Freiburg, Germany 2011.

A Flexible, Simulation-Based Approach to Sample Size Estimation for Prediction Modelling and Machine Learning: The pmsims R Package

13:45-14:00

Presenter: Oyebayo Ridwan Olaniran, Ewan Carr

Authors: Oyebayo Ridwan Olaniran [1]; Diana Shamsutdinova [2]; Sarah Markham [1]; Daniel Stahl [1]; Gordon Forbes [2]; Ewan Carr [1]

Affiliations: [1] King's College London, London; [2] King's College London, London

Introduction

Clinical prediction models are increasingly central to modern healthcare, yet most are developed using samples that are too small and without formal sample size justification, risking overfitting and poor generalisability. Existing tools rely on closed-form methods or simple simulation frameworks restricted to logistic, linear, and Cox regression models. No accessible software supports sample size estimation for machine learning or complex data structures.

Methods

We present pmsims, an open-source R package for simulation-based minimum sample size estimation in clinical prediction modelling. Unlike existing approaches that use a mean criterion, pmsims adopts a more conservative assurance criterion, targeting sample sizes at which performance meets the target with high probability (≥80%). This accounts for between-sample variability and is especially important for machine learning models, which often show greater instability across samples. pmsims is a modular, model-agnostic framework supporting standard regression models (linear, logistic, and survival), penalised regression, Random Forest, and XGBoost, while allowing users to supply their own data generators, model-fitting functions, and performance metrics. It uses an adaptive Gaussian process (GP) search engine to target informative sample sizes, reducing computation compared with exhaustive simulation.

Results

A validation study showed that the adaptive GP engine reliably achieves pre-specified performance targets across binary, continuous, and survival outcomes, with results consistent with existing tools in standard settings and requiring fewer model evaluations than non-adaptive approaches [1]. This talk will present results extending these findings to machine learning models and flexible data-generating mechanisms that encode correlations, interactions, and nonlinearities, settings for which pmsims is uniquely designed, and for which existing tools offer no guidance.

Discussion

The adaptive GP-based search offers a compelling middle ground: the flexibility of simulation-based approaches with the efficiency of adaptive search. By supporting sample size estimation across a wide range of model classes and data settings, pmsims represents a meaningful step towards making principled sample size planning the norm rather than the exception in clinical prediction research, addressing a persistent gap in current practice.

[1] Olaniran OR, Shamsutdinova D, Markham S, Zimmer F, Stahl D, Forbes G, Carr E. Adaptive Gaussian process search for simulation-based sample size estimation in clinical prediction models. arXiv 2026. https://arxiv.org/abs/2603.23688

Clinical prediction models in rare cancers: using patient perspectives to inform sample size calculations for model development with survival data

14:00-14:15

Presenter: Laura Kirton

Authors: Laura Kirton; Kym Snell; Rebecca Whittle; Richard Riley

Affiliations: University of Birmingham, Birmingham

Introduction

Early findings from a systematic review of prognostic models in sarcomas, a group of rare cancers, indicate that sample size is rarely considered prospectively, resulting in models being developed with insufficient data. Minimum sample size guidance for developing a prediction model with a time-to-event outcome, proposed by Riley et al. [1], provides closed-form calculations targeting minimal overfitting, precise estimation of overall risk and small optimism in R2. However, in rare cancers these recommended sample sizes can be unattainable. Patient perspectives highlight that individualised prediction remains valuable despite the increased uncertainty and instability associated with smaller sample sizes [2].

Methods

The general sample size framework is a simulation-based approach [3]. Key steps include generating large synthetic development and target population datasets, specifying a reference model using key predictors, developing models across a range of candidate sample sizes, and evaluating performance using metrics such as calibration, discrimination, and clinical utility to quantify performance degradation. We apply the framework to survival outcomes using Ewing sarcoma, a rare cancer, as the clinical exemplar, and show that engagement with patient advocates is essential for defining acceptable performance targets and ensuring sample size recommendations reflect the priorities of the intended population. A simulation study is conducted to identify key drivers of the required sample size.

Results

In the rare cancer example, patient advocates were accepting of relaxed target criteria, including wider uncertainty intervals around individual risk estimates. Using the general sample size framework and clear communication tools, patient perspectives inform appropriate target measures, allowing pragmatic sample sizes to be used in rare cancer settings and ensuring models are developed with multi-stakeholder endorsement. The simulation study shows how changing the performance targets influences the required (minimum) sample sizes. Where criteria must be relaxed to achieve feasible sample sizes, the framework provides transparency around potential limitations, supporting informed decision making for model development.

Discussion

This work demonstrates how the general sample size framework adapts for rare cancers, ensuring the developed prognostic models remain relevant and usable despite limited sample sizes. Integrating patient perspectives throughout promotes transparency and supports informed, context-appropriate model development and application.

References

[1] doi:10.1002/sim.7992

[2] doi:10.1136/bmj-2024-080749

[3] doi:10.48550/arXiv.2504.18730

SurvivalGPU: Scalable Survival Analysis for R and Python

14:15-14:30

Presenter: Antoine Poirot-Bourdain

Authors: Antoine Poirot-Bourdain [1, 2]; Alexis van Straaten [2, 3]; Emmanuel Bacry [1]; Anne-Sophie Jannot [2, 4]; Jean Feydy [2]

Affiliations: [1] CEntre de REcherche en MAthématiques de la DÉcision CNRS-UMR 7534 Université Paris-Dauphine PSL, Paris; [2] HeKa Université Paris Cité, INRIA, INSERM, F-75015, Paris; [3]

Assistance Publique-Hôpitaux de Paris (AP-HP), Service d'informatique Médicale, Paris; [4] Banque Nationale de Données Maladies Rares, Direction des Services Numériques, AP-HP, Paris

Introduction

Applying the Cox model to large-scale databases, such as electronic health records or medico-administrative data systems, faces significant computational bottlenecks. It is even more the case when modelling time-varying exposure such as the Weighted Cumulative Exposure (WCE) or when bootstrapping or cross-validation is needed. We introduce survivalGPU, an R and Python package leveraging GPU acceleration to address these limitations.

Methods

The package survivalGPU is implemented in Python using PyTorch and KeOps, with an R interface via reticulate. The coxph and wce functions of survival and wce packages were called respectively coxphGPU and wceGPU in survivalGPU package. We performed unit tests to assess whether survivalGPU acts similarly to survival and WCE packages. We compared estimates provided by survivalGPU against the survival and WCE packages on simulated datasets, comparing estimates from coxphGPU and wceGPU respectively. Performance was benchmarked on datasets from 10,000 to 1,000,000 patients (Cox model) and 1,000 to 100,000 patients (WCE model). We also evaluated the performance of the models with 1,000 bootstraps. Computations were done using a NVIDIA RTX 3090 GPU and an Intel Xeon W-2255 CPU.

Results

Both coxphGPU and wceGPU produced estimates identical to their reference implementations, with relative differences below 10^{-5}, smaller than Python alternatives such as lifelines or statsmodels. For the Cox model, survivalGPU is 10-20× faster than the R survival package on large datasets, with the gap widening substantially when performing bootstraps: 1,000 replicates on 100,000 patients takes 4,269s using survival versus 280s using survivalGPU. Gains are even larger for the WCE model: estimations on 100,000 patients takes ~10,000s with WCE versus 20s (~500× faster), and 1,000 bootstraps reduce from an estimated 10^7s to 2,000s (~5,000× faster).

Discussion

survivalGPU provides substantial computational speedups over reference implementations while maintaining numerical equivalence, making survival analysis more tractable on large-scale datasets. GPU acceleration is particularly impactful for bootstrap-based inference, reducing computation times from days to minutes. The package is freely available in R and Python and designed to integrate into existing statistical pipelines. Future work will extend support to competing risks, multiple concurrent exposures, and regularization in high-dimensional settings.

Robust Bayesian Modelling for Dynamic Prediction of HIV/AIDS Progression Across Regions

14:30-14:45

Presenter: Zahra Seyedi Lahroudi

Authors: Zahra Seyedi Lahroudi; Taban Baghfalaki

Affiliations: University of Manchester, Manchester

Introduction

Understanding disease progression and predicting patient outcomes are critical challenges in HIV/AIDS research. Longitudinal biomarkers, such as CD4 counts, provide rich information on immune function over time, while survival outcomes capture mortality risk. Existing models often rely on normality assumptions and ignore geographic clustering, limiting their robustness and applicability. Heavy-tailed variation and spatial heterogeneity are common in clinical and epidemiological data, particularly when patients are clustered within administrative regions such as states or provinces. Robust statistical methods that integrate longitudinal trajectories, survival outcomes, and spatial dependence are essential for accurate dynamic risk prediction and informed clinical decision-making.

Methods

We propose a robust Bayesian spatial joint model that addresses both outlier-prone longitudinal data and regional dependence in survival outcomes. The longitudinal sub-model uses a t distribution for measurement errors and random effects, providing resilience against extreme values. The survival sub-model incorporates regional spatial frailty using an intrinsic conditional autoregressive (ICAR) prior, extended with the Besag-York-Mollié (BYM2) parameterization. Four configurations, assuming normal or t distributions for random effects and errors, are systematically compared. Fully Bayesian inference is performed using Hamiltonian Monte Carlo, enabling dynamic landmark-based conditional survival predictions, which update individualized risk estimates as new biomarker data become available.

Results

Simulation studies demonstrate that the model yields reliable parameter estimates, well-calibrated credible intervals, and superior predictive performance compared with models ignoring heavy-tailed effects or spatial dependence. Applied to AIDS data from 27 Brazilian states, including CD4 count measurements and mortality outcomes, the model captures regional mortality disparities, generates dynamic individualized survival predictions, and enables comparison of alternative model specifications. Ignoring heavy-tailed random effects or spatial structure results in biased inference and reduced predictive accuracy.

Discussion

The proposed framework is robust to non-normality and captures the interplay between individual-level biomarker trajectories and geographic context. By providing dynamic risk predictions, it supports real-time clinical decision-making and personalized prognosis in HIV/AIDS. This methodology is broadly applicable to spatially referenced epidemiological data, offering a principled approach for integrating longitudinal, survival, and spatial information.

Combining disease-specific and other-cause mortality hazards from registry and population data to predict life expectancy and treatment-related life gains

14:45-15:00

Presenter: Cláudia Cruz Oliveira

Authors: Cláudia Cruz Oliveira [1]; Daan Nieboer [2]; Toby Hackmann [3]; Liesbeth de Wreede [3]; Berdine Heesterman [4]; Anne Stiggelbout [3]; Hester Lingsma [1]; Monique J. Roobol [2]; David van Klaveren [1]

Affiliations: [1] Department of Public Health, Erasmus MC University Medical Center, Rotterdam; [2] Department of Urology, Erasmus MC University Medical Center, Rotterdam; [3] Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden; [4] Department of Research and Development, Netherlands Comprehensive Cancer Organization, Utrecht

Introduction

A prostate cancer (PCa) diagnosis may be considered clinically irrelevant when the risk of significant disease (csPCa) is low and curative treatment offers limited survival benefit due to competing mortality risks. A previously developed tool predicts life expectancy (LE) with and without csPCa, and life-years gained from treatment. This tool was developed on an outdated cohort and made a simplified assumption that the hazard ratio of prostate cancer treatment for overall mortality is constant, regardless of competing mortality risks. We aimed to redevelop this tool with updated LE predictions and more accurate predictions of the life-years gained.

Methods

We included patients from the Netherlands Comprehensive Cancer Organization (IKNL) diagnosed from 2013 to 2024 (n=3672) and from the European Randomized Study of Screening for Prostate Cancer (ERSPC) diagnosed in Rotterdam, the Netherlands, between 1994 and 2000 (PCSM) (n=442). We estimated age-specific PCSM hazards from ERSPC data, using a flexible parametric competing risk model. We derived age- and CCI-specific other-cause mortality hazards by subtracting these PCSM hazard estimates from overall mortality hazard estimates based on 12 years of follow-up in the IKNL data. We extrapolated other-cause mortality hazards beyond 12 years based on Dutch life tables to obtain age- and CCI-specific life expectancy estimates conditional on active treatment. Based on the treatment hazard ratio for PCSM found in the PIVOT-trial, we obtained - per age and CCI - the life expectancy without treatment and the life years gained from treatment.

Results

The new model led to higher LE predictions, especially in younger patients without comorbidities. The new predictions of life-years gained were higher than the previous estimates for low mortality risk groups and lower for high mortality risk groups, better reflecting the attenuative effect of competing mortality risk for patients at older age with comorbidities.

Discussion

We redeveloped a shared decision-making tool that better predicts life expectancy and potential life-year gains from treatment. This tool supports clinicians and patients in making informed decisions about whether to proceed with a prostate biopsy.

15:30–17:00

Clinical Imaging

Mon, 28.09.26, 15:30–17:00, Room: K2

Chair: Heinz Handels

SSIM-Guided Disentangled Representation Learning for Scanner-Agnostic MRI Harmonization

15:30-15:45

Presenter: Luca Caldera

Authors: Luca Caldera [1]; Lara Cavinato [1]; Francesca Ieva [2, 3]

Affiliations: [1] MOX lab, Department of Mathematics, Politecnico di Milano, Milano; [2] MOX lab, Department of Mathematics, Politecnico di Milano, Milan; [3] HDS - Health Data Science Centre, Human Technopole, Milano

Introduction

Variability in MRI data due to differences in scanners, acquisition protocols, and imaging sites limits its reproducibility and generalizability in multicenter neuroimaging studies [1]. While feature-based harmonization methods adjust derived measures, they often fail to capture voxel-level variability. Image-based approaches address this issue but may struggle to preserve anatomical fidelity while enforcing appearance consistency [2]. We propose a novel 3D MRI harmonization framework that disentangles anatomical structure from scanner- and site-specific variations, improving cross-site consistency without compromising biological information.

Methods

We extend a domain translation architecture [3], originally designed for 2D natural images, to volumetric MRI data by introducing separate encoders for anatomical content and style, and a generator that recombines them. A key contribution is a differentiable loss function based on the Structural Similarity Index (SSIM) [4], enabling explicit control over luminance, contrast, and structural components. The model is trained on 3D T1-weighted MR images from multiple datasets (2,700+ scans, 88 sites, 19 scanners), including both healthy and clinical populations. Performance is evaluated in terms of anatomical preservation (SSIM), appearance consistency (Wasserstein distance and luminance similarity), and downstream predictive tasks.

Results

The proposed method achieves high anatomical fidelity, with SSIM values up to 0.975 between original and harmonized images. Inter-site variability is substantially reduced, as shown by a decrease in Wasserstein distance from 8.45 to 1.95 and improved luminance similarity (from 0.952 to 0.982). Harmonization also enhances downstream performance: brain age prediction error decreases from 4.08 to 2.8 years, while Alzheimer's disease classification improves (AUC from 0.86 to 0.90), with consistent gains in balanced accuracy and F1-score.

Discussion

The proposed framework effectively harmonizes MRI data across heterogeneous acquisition settings while preserving anatomically meaningful information. The SSIM-based loss enables a balance between structural fidelity and appearance alignment, overcoming limitations of existing methods. Improvements in downstream tasks highlight the practical impact of the proposed harmonization approach in both clinical and research applications. Overall, this approach provides a robust and generalizable solution for large-scale, multicenter neuroimaging studies.

References

1. Shinohara, R. T., Oh, J., Nair, G., Calabresi, P. A., Davatzikos, C., Doshi, J.,... & Bakshi, R. (2017). Volumetric analysis from a harmonized multisite brain MRI study of a single subject with multiple sclerosis. American Journal of Neuroradiology, 38(8), 1501-1509.

2. Hu, F., Chen, A. A., Horng, H., Bashyam, V., Davatzikos, C., Alexander-Bloch, A.,... & Shinohara, R. T. (2023). Image harmonization: A review of statistical and deep learning methods for removing batch effects and evaluation metrics for effective harmonization. NeuroImage, 274, 120125.

3. Li, X., Hu, J., Zhang, S., Hong, X., Ye, Q., Wu, C., & Ji, R. (2019). Attribute guided unpaired image-to-image translation with semi-supervised learning. arXiv preprint arXiv:1904.12428.

4. Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4), 600-612.

Ensemble Multiblock Scoring for Robust Subtype Identification in Multimodal Neuroimaging

15:45-16:00

Presenter/Author: Atsushi Kawaguchi

Affiliations: Saga University, Saga

Introduction

Multimodal neuroimaging provides rich phenotypic information relevant for characterizing disease heterogeneity, yet voxel-wise data remain extremely high-dimensional, leading to instability in conventional unsupervised learning. To address these limitations, we introduce an ensemble clustering

multiblock sparse multivariable analysis (ECMSMA) framework that integrates hierarchical dimension reduction with consensus clustering to derive statistically robust and clinically interpretable subgroups.

Methods

We analyzed baseline multimodal MRI data from OASIS-3 participants who were cognitively normal at baseline (CDR = 0 age ≥ 70 years). Six imaging modalities (T1w, T2w, TSE, T2*w, minIP, SWI) were spatially normalized, and voxel-level measurements were reduced using spherical basis functions to incorporate spatial correlation. At the block level, modality-specific principal component scores were derived via multiblock sparse multivariate analysis (MSMA), yielding multiple components per modality. These block-level scores were concatenated to form a super-dataset, transformed to non-negative representations, and clustered using non-negative matrix factorization (NMF), which also produced multiple candidate partitions. For each subject, clustering solutions across components were integrated using ensemble methods (majority voting, k-modes, CSPA, LCE, LCA). Component selection was guided by internal validity indices (e.g., Calinski-Harabasz, Dunn, Silhouette), and external validity was evaluated by concordance with longitudinal clinical trajectories defined by whether cognitive status remained stable or worsened based on post-baseline CDR changes.

Results

Internal validity metrics varied substantially across components and did not reliably reflect clinically meaningful structure. In contrast, ensemble clustering consistently improved external validity relative to any single-component solution, with majority voting over top-ranked components yielding the strongest agreement with longitudinal outcomes. The derived subgroups showed coherent modality-level contributions, with T1w exerting the greatest influence, and highlighted anatomically plausible regions, most notably hippocampus and amygdala, that align with early neurodegenerative processes. Bootstrap resampling (10,000 replicates) confirmed the robustness and reproducibility of ensemble-derived clusters.

Discussion

Hierarchical multiblock scoring combined with ensemble clustering provides a statistically principled and clinically interpretable strategy for subtype discovery in high-dimensional multimodal neuroimaging. By stabilizing clustering solutions and recovering biologically meaningful patterns, the ECMSMA framework may support early risk stratification and facilitate the development of data-driven biomarkers for preclinical neurodegeneration.

When Does Clinical Fusion Add Value Beyond Imaging in Multimodal Stroke Prognosis?

16:00-16:15

Presenter/Author: Vanya Valindria

Affiliations: Monash University, Jakarta

Introduction

Early prediction of neurological recovery after acute ischemic stroke guides treatment decisions and resource allocation in emergency settings. While 24-hour NIHSS change (ΔNIHSS-24h) predicts outcomes more reliably than baseline scores alone, existing models show inconsistent performance across different imaging contexts and backbone architectures. In resource-limited settings where only non-contrast CT (NCCT) and basic clinical data are available, it remains unclear when multimodal fusion provides meaningful improvements over image-only models. We systematically evaluate the conditions under which integrating clinical variables with NCCT imaging enhances prediction, addressing critical gaps in understanding backbone-fusion interactions and establishing practical decision thresholds for clinical deployment.

Methods

We developed CTRTab, a multimodal framework integrating NCCT, radiomics, and clinical data using frozen TabPFN encoders that natively handle missing values. We evaluated five pretrained vision backbones (ResNet18/50, DenseNet121, ViT-B/16, Swin-T) across four slice-selection policies (middle, most-infarct, lesion-only, cropped-lesion) and four fusion operators (concatenation, sum, max, feature-wise modulation) on two cohorts: Local UI dataset (n=81, improvement=ΔNIHSS≥2) and ISLES 2024 (n=79, improvement=>50% reduction). Performance was evaluated using macro F1 with stratified 5-fold cross-validation. Clinical-only baselines were evaluated to establish lower bounds and assess imaging necessity.

Results

Backbone and slice policy dominated performance: most-infarct slices achieved highest F1 scores (up to 87.3% with Swin-T on ISLES), with middle slices as robust alternatives. Lesion-only views degraded CNN performance while transformers showed resilience. Fusion benefits were backbone-dependent: CNNs gained substantial improvement on informative slices, while strong transformers on optimal slices showed minimal gains. However, under suboptimal conditions (middle/cropped slices), even transformers improved significantly (~22% with fusion). Clinical-only baselines were underperforming multimodal models, confirming imaging necessity.

Discussion

Our findings establish a fusion framework: (1) With good lesion visibility and segmentation tools, Swin-T image-only models achieve optimal performance (87.3% F1) (2) In resource-limited settings or with uncertain slice selection, lightweight CNNs with fusion offer comparable performance (~81% F1) at lower computational cost and greater robustness (3) Fusion consistently stabilizes performance across variable imaging conditions. TabPFN's native missing-value handling enables real-world deployment without imputation bias. Future prospective validation should include confidence intervals and calibration analysis to support clinical decision thresholds.

Survival Prediction of TAVI Patients Using End-to-End Deep Image Regression on Whole-Body CT

16:15-16:30

Presenter: Alexia Rizoudis

Authors: Dominik Müller [1]; Alexia Rizoudis [1]; Charlotte Lind [2]; Frank Kramer [3]; Jakob Christoph Voran [2]; Hannes Ulrich [4]

Affiliations: [1] Universität Augsburg, Bayern, Deutschland, Augsburg; [2] Department of Cardiology, Kiel University and University Hospital Schleswig-Holstein, Kiel, Germany, Kiel; [3] Lehrstuhl für IT-Infrastrukturen für die Translationale Medizinische Forschung, Universität Augsburg, Augsburg; [4] Institute for Medical Informatics and Statistics, Kiel University and University Hospital Schleswig-Holstein, Kiel

Introduction

Transcatheter aortic valve implantation (TAVI) is a standard treatment for patients with severe aortic stenosis. However, predicting long-term survival outcomes remains challenging, as traditional risk scores such as EuroSCORE II and STS-PROM exhibit poor calibration for TAVI cohorts.

Methods

In this work, we developed an end-to-end deep learning pipeline to predict overall survival of TAVI patients directly from pre-procedural whole-body CT scans. The pipeline uses a 3D DenseNet121 architecture within the AUCMEDI framework, trained as a regression model to predict survival in days.

Results

Using 5-fold cross-validation on 317 TAVI patients, the model achieved an AUC of 0.787, a specificity of 72.1%, and a sensitivity of 85.3% for identifying short-term survival patients (<12 months) in a binary evaluation.

Discussion

Notably, this image-only model achieves performance comparable to state-of-the-art multimodal machine learning models that require extensive clinical data. These results demonstrate that end-to-end deep learning on routinely acquired CT scans can effectively identify high-risk patients, offering potential as a clinical decision support tool for pre-operative TAVI assessment.

The full paper of this abstract will be published in the GMDS series "Studies in Health Technology and Informatics" with doi: 10.3233/SHTI261005, URL:https://dx.doi.org/10.3233/SHTI261005

Interpolation Methods for MediaPipe Hand Tracking Gaps A Pose-Based Approach

16:30-16:45

Presenter: Solveig Najork

Authors: Solveig Najork [1]; Michael Weber [1]; Gerrit Bücken [2]; Marco Mass [1]; Lisa Kunkel genannt Bode [3, 4]; Philipp Koch [5]; Marcin Grzegorzek [3, 1]

Affiliations: [1] German Research Center for Artificial Intelligence (DFKI), Lübeck; [2] Fraunhofer Research Institution for Individualized Medical Technology and Engineering (IMTE), Lübeck; [3] University of Luebeck, Lübeck; [4] University Hospitals Schleswig-Holstein, Lübeck; [5] Charité-University Medicine Berlin, Berlin

Introduction

Kinematics are a relevant measure of movement quality, especially for physically restricted patients. Therefore, consistent landmark detection is necessary. Automated landmarks detection already has existing solutions, but a problem remains: missing frames, where landmarks cannot be automatically detected.

Methods

This paper evaluates three interpolation strategies to overcome this issue and thus enable kinematic analysis. These strategies are linear interpolation, the Kalman filter, and the developed pose-based approach. Recordings were obtained using the marker-based VICON system. Participants performed three tasks inspired by a standard stroke assessment. This setup allows the precision of marker-less detection and interpolation to be evaluated.

Results

In comparison with Percentage Root Mean Square Error (PRMSE) and the Procrustes mean disparity, linear interpolation yields the best results. It achieved a mean PRMSE of 36.5 and a mean disparity of 0.266. The pose-based approach, presented here, follows closely with a mean PRMSE of 36.6 and a mean disparity of 0.273.

Discussion

In subsequent optical evaluations, the pose-based approach demonstrated superior performance. Our results demonstrate that a reliable estimate with consistent landmarks is possible, thus enabling kinematic analysis.

The full paper of this abstract will be published in the GMDS series "Studies in Health Technology and Informatics" with doi: 10.3233/SHTI261006, URL:https://dx.doi.org/10.3233/SHTI261006

Performance uncertainty in medical image analysis: a large-scale investigation of confidence intervals

16:45-17:00

Presenter: Pascaline André

Authors: Pascaline André; Charles Heitz

Affiliations: Paris Brain Institute, Paris

Introduction

Performance uncertainty quantification is essential for reliable validation and eventual clinical translation of artificial intelligence (AI) developed tools. Confidence intervals (CIs) play a central role in this process by indicating the precision of a reported performance estimate. Yet, due to the limited work examining CI behavior in medical imaging, the community remains largely unaware of the diversity of CI methods and how they behave in specific settings. The purpose of this study is to close this gap.

Methods

To this end, we conducted a large-scale empirical analysis across a total of 24 segmentation and classification tasks, using 19 trained models per task group, a broad spectrum of commonly used performance metrics, multiple aggregation strategies, and several widely adopted CI methods. A description of the trained models and tasks can be found in [1] for segmentation and [2] for classification. Reliability (coverage) and precision (width) of each CI method were estimated across all settings to characterize their dependence on study characteristics.

Results

Our analysis revealed five principal findings: 1) the sample size required for reliable CIs varies from a few dozens to several thousands of cases depending on study parameters 2) CI behavior is strongly affected by the choice of performance metric 3) aggregation strategy substantially influences the reliability of CIs, e.g. they require more observations for macro than for micro 4) the machine learning problem (segmentation versus classification) modulates these effects 5) different CI methods are not equally reliable and precise depending on the use case. For instance, the BCa bootstrap method fails for CIs of the median, whereas the percentile bootstrap is the most robust bootstrap method. We also noticed that some parametric approaches outperform bootstrap for very small samples but this advantage vanishes rapidly.

Discussion

These results form key components for the development of future guidelines on reporting performance uncertainty in medical imaging AI.

[1] Antonelli, M. et al. (2022). The medical segmentation decathlon. Nature communications, 13(1):4128.

[2] Godau, P. et al. (2024). Beyond knowledge silos: Task finger-printing for democratization of medical imaging ai. arXiv preprint arXiv:2412.08763.

GMDS Awards

Mon, 28.09.26, 15:30–17:00, Room: K3

Chair: André Scherag, Hans-Ulrich Prokosch

Master's thesis "The effect of smoking on the composition of the anterior nares microbiome"

Presenter/Author: Angelika Wassmuth

Introduction There is a debate if smoking can cause or exacerbate respiratory diseases or infections by alteration of the respiratory tract microbiome. Previous research suggests that smokers have a lower microbiome richness and a higher abundance of anaerobic pathobionts than never smokers. However, evidence from the anterior nares' microbiome is limited. Additionally, there is a considerable degree of heterogeneity in analysis strategies for microbiome data. This thesis aimed to develop a reproducible, R-based analysis pipeline and to apply this to characterise the effect of smoking on the anterior nares' microbiome in the German National Cohort (NAKO).

Methods This project used data from the baseline examination of NAKO, comprising 1,950 participants (16.8% current smokers, 36.6% former smokers) with 16S rRNA sequencing (V3-V4 regions) of nasal swabs from the anterior nares. A comprehensive, R-based analysis pipeline was developed to ensure reproducibility and robustness when facing microbiome data challenges. The pipeline consists of seven modules: quality control, alpha diversity, beta diversity, Latent Dirichlet Allocation (LDA) clustering, and differential abundance analysis (DAA) using Wilcoxon rank sum test, MaAsLin3, and a random forest model.

Results Current smokers exhibited higher species evenness (Shannon and Inverse Simpson indices) than never smokers. PERMANOVA indicated that smoking status explained 0.4-1.1% of variance. Thirty-six genera were identified as being differentially abundant between smoking statuses, with eight being identified by all three DAA methods. Anaerococcus, Finegoldia, and Peptoniphilus were identified as being differentially abundant in current and former smokers. The optimal LDA model contained six topics. The distribution of topics across smoking status groups was consistent with the DAA results.

Discussion Smoking was associated with higher evenness and enrichment of anaerobic pathobionts in the anterior nares. The project's main strengths are its large, population-based sample and the use of a reproducible, R-based pipeline that enables comprehensive microbiome profiling. However, the findings are limited in their generalisability, as the NAKO cohort under-represents certain demographic groups, and the taxonomic resolution is restricted to the genus level. Future longitudinal analyses are needed to assess causal pathways and recovery dynamics following cessation.

Sample Size Recalculation in a Hybrid Adaptive Clinical Trial Design for Time-to-Event Endpoints

Presenter/Author: Anna Linger

Usually, hypothesis testing is only conducted after all patients have been recruited and observed in a clinical trial. For trials with a time-to-event endpoint, the amount of information available at the analysis time point does not only depend on the number of recruited patients but also on the number of observed events, which is random for a fixed time point. This leads often to long trial durations. Thus, it might be desirable to evaluate the accumulated information during the ongoing study and have the possibility to adapt the sample size at interim analyses to address planning uncertainties. Adaptive group-sequential designs address those challenges by providing the option for an early trial stop and adapt the sample size based on the observed data at the interim analysis while maintaining the trial integrity and type I error rate. Agencies usually prefer blinded interim analyses to avoid potential bias. However, blinding cannot always be maintained, especially if there is a high interest in an early trial stop or a high insecurity of the underlying effect size. Hence, so called hybrid group-sequential designs were recently proposed, which aim to combine the advantages of an always blinded single-stage design and an adaptive group-sequential design with unblinded interim analyses. Those designs use a blinded estimator of the number of events that occurred until the interim analysis to derive a decision whether to unblind the trial or to continue it as it is. However, the unblinded interim analysis only allowed for early stopping but not for a sample size adaptation. In this work, we add the option of recalculating the sample size in case the trial is unblinded during the interim analysis and continued to the second stage, i.e. a hybrid adaptive design with sample size recalculation. This extension is based on the inverse normal combination test and the independent increment structure of the log-rank test statistic while controlling the overall type I error rate. The hybrid adaptive design is compared to the single-stage and classical adaptive design with respect to type I error, power, bias, unblinding probability, expected sample size and study duration in comprehensive simulation studies.

Adaptive study designs with restricted mean survival times and sample size adaptations

Presenter/Author: Carolin Herrmann

Affiliations: Heinrich-Heine-Universität Düsseldorf, Düsseldorf

When planning a clinical trial, it can sometimes be useful to apply a design with an interim analysis. It makes it possible to stop the trial early and to potentially also make design adaptations without undermining the study validity and integrity. Consequently, many statistical methods for such adaptive designs have already been developed, yet not for all application scenarios. In a statistical consultation, a framework for sample size adaptations with the restricted mean survival time was preferred for planning a specific clinical trial. However, such a design was not available at that timepoint.

Therefore, we developed the corresponding statistical methods and conducted an intensive simulation study. We compared the design with other commonly chosen test strategies in the survival setting. In this talk, we present the results of this work, as recently published in Statistics in Medicine. Moreover, we discuss not only the new method's performance but also the choice of an appropriate estimand for the clinical trial example.

This work has been published in: Herrmann, C., & Blanche, P. (2026). Sample Size Recalculation in Adaptive Group Sequential Study Designs for Comparing Restricted Mean Survival Times. Statistics in Medicine, 45(6-7), doi: 10.1002/sim.70490.

Design and Development of an Assistant-Supported Validation Process for EDC Systems as Software-as-a-Service in Clinical Research

Presenter/Author: Celine Magnus

Computer System Validation (CSV) is an integral component of the regulatory landscape in the pharmaceutical and medical sector and is generally regarded as complex and time-con suming. This includes the validation effort necessary for applications used for (electronic) data capture (EDC) in clinical trials. To counter such criticism, an assistant-supported validation approach is required. The goal of the presented thesis is to develop and partially prototype such an approach based on regulatory requirements and the results of a literature review in the context of "Daisy", a Software-as-a Service (SaaS) EDC system currently under development by the involved industry partner.

The materials employed and the methods applied in the thesis include the consideration of relevant regulatory regulations and industry standards as well as standards of the industry partner. Furthermore, the methodology for a systematic literature and market review to evaluate the current state of research and industry is introduced as well as the criteria-based development of a concept for the assisted validation process. With regard to the development of the prototype, tools and methods such as SCRUM, Laravel and Vue were specified by the industry partner. Unscripted tests and a user survey containing the System Usability Scale (SUS) questionnaire are used in the evaluation of the prototype.

The results of the regulatory requirements analysis (in the form of a requirement catalogue) as well as literature and market review provide a good foundation for the development of a concept for an assisted validation process in the context of "Daisy", consisting of several approaches for the different subprocesses. The approach for the validation plan's subprocess is implemented as a wizard prototype and subsequently evaluated. The evaluation identifies strengths and potential for improvement for the wizard and confirms the fulfilment of prerequisites for regulatory compliance.

The catalogue of requirements as well as literature and market review provide a solid base for the development of the concept which can be further refined. The wizard prototype is evaluated as highly promising and will be the object of further development for Daisy. Therefore, an innovative approach to an assistant-supported and automated validation process has been ultimately developed, which builds the foundation for future solutions.

A Simulation-Based Comparison of Winner's Curse Correction Methods in the Context of Multi-Group Animal Experiments

Presenter/Author: Henrike Solveen

Affiliations: Institute for Medical Biometry, Informatics and Epidemiology University of Bonn, University Hospital Bonn, Bonn; Koblenz University of Applied Sciences, RheinAhrCampus Remagen., Remagen

Exploratory animal experiments are essential for evaluating new drugs and interventions. Due to ethical and logistical constraints, they are typically conducted with small sample sizes. Moreover, these experiments often involve the comparison of multiple groups. As a consequence, a large number of statistical tests are performed, each based on only a few animals. In practice, only statistically significant findings are typically reported. The combination of small sample sizes, small effect sizes, multiple comparisons, and significance thresholds leads to systematically inflated effect size estimates, a phenomenon commonly referred to as the winner's curse. This systematic overestimation of effect sizes affects the planning and conduct of follow-up studies, as well as the translation to clinical trials. In addition to wasting resources, it also raises ethical concerns, as animal lives may be spent on findings that fail to replicate. Currently, no established methodology exists for mitigating overestimated effect sizes in the context of exploratory animal experiments. Therefore, the aim of this work was to explore, apply, and evaluate statistical approaches for correcting the winner's curse that were originally developed in other research fields. Following a systematic literature review to identify candidate methods, three approaches originating from genomic research and one newly developed method for correcting effect size overestimation in animal experiments were evaluated in a comprehensive simulation study. The simulated scenarios reflected typical preclinical study designs and varied with respect to the number of groups, sample size per group, and underlying effect sizes. The results show that most of the methods can mitigate the winner's curse, although their performance depends on the simulation scenario. Moreover, they highlight that the assumptions required for the methods identified from the literature are not ideally met in practice, particularly because of the small sample sizes inherent to animal experiments. These findings underline both the potential and the limitations of current approaches and emphasize the need to develop methods specifically tailored to preclinical research.

Unsupervised Learning of Cell Type Distributions in Spatial Transcriptomics Data via Spatially-informed Non-Negative Matrix Factorization

Presenter: Malte Mensching-Buhr

Authors: Michael Altenbuchinger

Affiliations: Universitätsmedizin Göttingen, Göttingen

Low-resolution sequencing-based spatial transcriptomics platforms produce bulk transcriptomics profiles per analyzed spot. These gene expression values consist of convoluted contributions of different cell types in heterogeneous samples. Several supervised learning approaches for the cell-type deconvolution of bulk profiles have been proposed and established. However, they neglect the spatial context of the data and rely on suitable training data from single-cell experiments. Here, we introduce an unsupervised learning approach for the cell-type deconvolution of spatial transcriptomics data called Spatially-informed Non-negative Matrix Factorization (SNMF), which combines Non-negative Matrix Factorization with Gaussian Kernel Interpolation to utilize the data's underlying spatial context. It can predict globally distributed cell types directly from the expression data matrix cost-efficiently and provides better estimates than classical clustering approaches such as Principal Component Analysis. It comes with an efficient GPU-accelerated implementation that can be integratedinto supervised learning approaches as well.

Investigator-initiated versus industry-sponsored trials - visibility and relevance of randomized controlled trials in clinical practice guidelines (IMPACT)

Presenter/Author: Manuel Hecht

Background: Randomized controlled trials (RCTs) are of outstanding importance for evidence-based medicine, particularly for assessing treatment effects. Despite their significance, the enormous resource investment, and the ethical obligation toward participants, RCT results are often inadequately published or fail to achieve sufficient visibility in clinical practice. This work therefore investigated the influence of German investigator-initiated trials (IITs) on clinical practice guidelines (CPGs) compared to international IITs, German industry-sponsored trials (ISTs), and international ISTs.

Methods: This work builds on the previously conducted IMPACT project, a cohort study of 691 confirmatory therapeutic multicenter RCTs, for which a range of characteristics were extracted from trial registries and associated publications, systematic reviews, and CPGs were identified in a standardized manner. In the present study, all associated CPGs were acquired in full text, guideline characteristics were extracted in a standardized fashion, and all citations of IMPACT trial references within guidelines were identified and classified. This enabled quantification of guideline impact per RCT and analysis of time-to-guideline-impact.

Results: Overall, 177 RCTs were cited in 405 guidelines, though only 153 RCTs achieved actual impact on 331 guidelines. German IITs were non-inferior to the other cohorts regarding guideline impact, a finding confirmed in sensitivity analyses incorporating additional explanatory variables and more robust outcomes. Regarding time-to-guideline-impact, German IITs were likewise non-inferior in both non-parametric and parametric analyses accounting for potential confounders. In an additional comparison between all cohorts using analogous methods, international ISTs were inferior to the other cohorts with respect to guideline impact, number of guidelines influenced, and time-to-guideline-impact - a result supported by post-hoc analysis. Furthermore, a positive association between trial

sample size and both guideline impact and time-to-guideline-impact was observed. Conclusion: This work demonstrates the non-inferiority of German publicly funded RCTs (conducted 2005-2016) regarding guideline impact, alongside evidence of inferiority among international industry-sponsored RCTs - consistent with previously described publication disparities between ISTs and IITs. Beyond these findings, this study shows that a substantial proportion of published RCTs are not incorporated into evidence syntheses, underscoring the need for regulatory efforts to ensure the visibility of all future RCTs.

Capturing dynamics under misspecification and potential transformation

Presenter/Author: Maren Hackenberg

Affiliations: European Molecular Biology Laboratory (EMBL), Heidelberg

Modeling dynamics is central to understanding complex processes in biomedicine, such as disease progression, cellular differentiation, or population dynamics. However, practical challenges, such as limited data, insufficient knowledge about the assumed dynamics, heterogeneity and high dimensionality often require to work with misspecified models and under transformation. In my thesis, I explore how dynamics can still be captured under such constraints by developing novel methodological approaches. I combine a theoretical characterization of prototypical classes of modeling problems in the context of misspecification and transformation with an empirical solution of the estimation task that is feasible in practice. The developed approaches are grounded in specific applications, including ICU utilization forecasting, time-series single-cell gene expression, and rare disease progression. At the same time, they serve as prototypical templates for broader dynamic modeling problem classes. Across these applications, I show how deliberate model misspecification and transformation via dimensionality reduction can make inference feasible while preserving meaningful dynamic structure. It further highlights the trade-off between model complexity or resolution and model estimability given data limitations. Dimensionality reduction is investigated both as an implicit model for capturing dynamics in a low-dimensional space, and as a prerequisite for incorporating an explicit dynamic model in such a space. Jointly estimating dynamic modeling and dimensionality reduction via differentiable programming offers a strategy to shape and evaluate a low-dimensional representation with respect to its ability to preserve temporal structure. Taken together, the thesis shows how mathematical and statistical modeling can be effectively combined with data-driven approaches to address dynamic modeling problems under misspecification and transformation. It contributes a coherent and structured framework for addressing broader classes of dynamic modeling challenges beyond the specific applications studied, and offers novel, flexible solutions across different domains.

GenRare CDM: A Standardized and Reusable Framework for Integrating Genomic, Phenotypic, and Clinical Rare Disease Data for Multi-Site AI-Driven Research

Presenter/Author: Najia Ahmadi

Rare diseases present major challenges for biomedical research and clinical care. Although each individual rare disease affects only a small number of patients, collectively they impact millions of people worldwide, with approximately 80% having a genetic basis. Clinical management is hindered by disease heterogeneity, delayed diagnosis, fragmented documentation, and the lack of large, harmonized datasets that integrate clinical, phenotypic, and genomic information. While advances in sequencing technologies and computational methods have created new opportunities for precision medicine, their full potential depends on transforming heterogeneous data into standardized, semantically consistent, and computationally reusable formats.

The "GenRare CDM: A Standardized and Reusable Framework for Integrating Genomic, Phenotypic, and Clinical Rare Disease Data for Multi-Site AI-Driven Research" dissertation addresses this challenge through the development and evaluation of the GenRare CDM, a structured seven-step pipeline for integrating genomic, phenotypic, and clinical rare disease data using the OMOP Common Data Model (OMOP CDM) as the central standardization framework. The primary objective is to enable interoperability of rare disease data across institutions and national borders while supporting downstream patient-level prediction and AI-driven analyses. In contrast to previous disease-specific solutions, this work provides a generalizable methodological framework that can be applied across multiple disease domains and healthcare settings.

A central contribution of this dissertation is the integration of standard OMOP clinical tables with genomic and phenotypic extensions. The resulting data model comprises seven core domains: Person, Diagnosis, Procedure, Medication, Laboratory Values, Genotypes, and Phenotypes. Clinical variables are standardized using established terminologies, including SNOMED CT, ICD-10-GM, LOINC, and the Human Phenotype Ontology (HPO). Because HPO is not natively supported within OMOP, this work introduces an integration strategy based on SOURCE_TO_CONCEPT mappings. This approach preserves OMOP compatibility while enabling standardized representation and computational use of phenotypic information, thereby addressing a key challenge in semantic interoperability.

The proposed pipeline was implemented and evaluated across four clinical domains, namely endocrinology, gastroenterology, pneumology, and hematology, using real-world datasets from Germany and the United States. Overall, the framework integrated data from 61,697 patients, including acute myeloid leukemia (AML) cohorts comprising 3,298 patients. The evaluation demonstrated that most of genomic variables could be mapped directly to standardized concepts, while automated ETL workflows reduced manual mapping effort. Furthermore, predictive models developed for AML achieved AUROC values exceeding 0.82 and successfully identified established genomic biomarkers, including NPM1 and FLT3-ITD, together with clinical variables such as white blood cell count, as important predictors of patient outcomes.

By combining conceptual data modeling, semantic harmonization, technical implementation, and analytical validation, this dissertation addresses critical barriers to translational rare disease research. It demonstrates the value of standardized, AI-ready data infrastructures for integrating heterogeneous

biomedical data and provides a scalable foundation for future multi-institutional studies, precision medicine, and AI-driven rare disease research.

Introduction of the electronic health professional card (eHBA) at a university hospital, using Hannover Medical School (MHH) as an example - Development of an Ideal Process

Presenter/Author: Selin Gürbulak

The digital transformation of the German healthcare system increasingly relies on applications of the telematics infrastructure (TI), including the electronic patient record (ePA) and the electronic prescription (eRezept). Healthcare professionals are required to use the electronic health professional card (eHBA) to access TI applications, which serves as a secure digital identity and enables qualified electronic signatures. As university hospitals are legally required to connect to the TI and operate its applications, an efficient and standardized implementation process for the eHBA has become essential. This paper addresses the lack of recommendations for implementing the eHBA in a structured manner in university hospitals in Germany. Hannover Medical School (MHH) was selected as an example to analyze the existing implementation process and identify opportunities for improvement. A mixed-methods approach was applied, including a literature review, qualitative expert interviews with MHH IT staff and physicians, an online survey among German university hospitals and Business Process Model and Notation (BPMN 2.0) process modelling.

Findings revealed that organizational obstacles, rather than technical limitations, represent the main barriers to successful implementation. In particular, unclear responsibilities and communication, bureaucratic application procedures and limited user acceptance were identified as key difficulties. Based on these findings and a comparison with implementation approaches at other university hospitals in Germany, an ideal process was developed, emphasizing centralized coordination and responsibilities, structured communication and user support.

The proposed ideal process provides a practical and reproducible recommendation for implementing the eHBA in university hospitals in a resource-efficient and cost-effective manner. Although this paper focuses on physicians, the proposed process can be adapted to other healthcare professional groups eligible for an eHBA through organizational adjustments.

Adaptive Trial Inference and Error Control

Mon, 28.09.26, 15:30–17:00, Room: K5

Chair: Lukas Andreas Widmer

Multiple type I error concepts for clinical trials with overlapping populations

15:30-15:45

Presenter: Remi Luschei

Authors: Remi Luschei; Werner Brannath

Affiliations: University of Bremen, Bremen

Introduction

The population-wise error rate (PWER) was introduced as a less conservative alternative to the family-wise error rate (FWER) for clinical trials with multiple, overlapping patient populations. These trials are particularly relevant in personalized medicine, which aims to find therapies tailored to different patient subgroups. By controlling an average multiple type I error probability over all population strata, the PWER can substantially improve statistical power. However, one disadvantage of this concept is that the error probability for a given population can strongly depend on the presence or absence of other populations included in the analysis.

Methods

To address this issue, we propose two modifications of the PWER that enforce individual error control either for all target populations, or for all possible unions of target populations. We call these approaches the PWER over the populations (PWER-P) and the PWER over population unions (PWER-U).

Results

We apply the new error rates to various study settings, including overlapping, nested and disjoint populations. In many situations, they provide a good compromise between PWER and FWER control, enabling stricter type I error control than the PWER, but still higher power than the FWER. The PWER-P and PWER-U can be controlled similarly to the PWER. In general, however, it is not possible to exhaust all population-wise errors at the chosen significance level, as these may be interdependent due to population overlaps. We also show that estimating the strata-wise prevalences, which is often necessary in practice, does not substantially inflate the true PWER-P and PWER-U.

Discussion

The PWER-P and PWER-U help to prevent a certain risk of misuse that exists with the PWER: disadvantaging populations at type I error control through the unequal allocation of error levels. However, we still recommend to use the PWER in certain situations, such as with nested populations, where both the PWER-P and PWER-U coincide with the FWER.

W. Brannath, C. Hillner, K. Rohmeyer. The population-wise error rate for clinical trials with overlapping populations. Statistical Methods in Medical Research, 32(2):334-352, 2023.

Weighted Two-Part tests for the Analysis of Zero-Inflated Data

15:45-16:00

Presenter: Fabrice Lotola Mougeni

Authors: Fabrice Lotola Mougeni; Martin Posch; Sonja Zehetmayer

Affiliations: Medical University of Vienna, Center for Medical Data Science, Institute of Medical Statistics, Vienna

Introduction

In clinical trials for treatments of worm infections, microfilaria counts are frequently measured as primary outcome. The distribution of these counts typically follows a mixture distribution, with a point mass at zero. For fixed sample clinical trial designs, comparing a treatment to a control, non-parametric methods like the Mann-Whitney U-test are commonly used to analyze such count data. For two-sided tests, in addition two-part tests have been proposed. These tests aim to improve the power by separately comparing the proportion of zeros and the locations of the distributions of the positive counts. The aim of this work is to generalize these two-part tests to the one-sided setting and make them more robust when the proportion of zeros is unknown.

Methods

We first extend previously proposed two-sided two-part tests to the one-sided setting. The considered tests are based on two one-sided tests, applied to the observed counts: a Mann-Whitney-U test, computed from observations with

positive counts, and a Fisher exact test, obtained by dichotomizing the count data. The p-values from these two tests are then combined with the inverse normal combination function to obtain an overall one-sided p-value. Furthermore, we propose

adaptive two-part tests where the weights in the inverse normal combination function are chosen in a data-dependent way based on the blinded outcome data. We prove that the resulting testing procedure controls the type I error rate.

Results

The simulation study confirmed the analytical results and the pro-posed one-sided two-part tests controlled the type I error rate when the weights were prespecified from blinded data. The weighted two-part tests showed higher power than the Mann-Whitney U-test and zero-inflated regression when the treatment effect was present mainly in the binary or in the positive-count part and the weights were chosen based on corresponding assumptions. When effects were present in both components, power was generally similar to that of the Mann-Whitney U-test, with performance depending on the proportion of zeros and the degree of dispersion.

Discussion

The proposed weighted one-sided two-part tests are a nonparametric approach for zero-inflated count outcomes that remains valid when the weights are chosen from blinded data.

A note on blinded continuous monitoring for continuous outcomes

16:00-16:15

Presenter: Long-Hao Xu

Authors: Long-Hao Xu; Tim Friede

Affiliations: University Medical Center Göttingen, Göttingen

Introduction

Continuous monitoring can improve efficiency in two-sample studies by allowing sample size updates as nuisance parameters, such as the outcome variance, are estimated during recruitment. In clinical trials, blinded monitoring is especially attractive because it preserves treatment concealment and reduces the risk of operational bias. However, although blinded continuous monitoring has been studied previously by simulation, its theoretical properties have not been fully established. This work addresses that gap for continuous outcomes.

Methods

The paper considers a two-group normal outcome setting and studies continuous monitoring rules based on variance estimation. The blinded procedure uses a pooled variance estimator that ignores treatment allocation, whereas the unblinded procedure uses the standard within-group pooled variance. The authors derive finite-sample and asymptotic properties of the blinded stopping rule, including termination, moment bounds, and limiting behavior under different regimes in which either the planning constant v or the variance sigma^2 becomes large. Simulation studies with 10,000 repetitions were conducted to assess finite-sample performance and compare blinded and unblinded monitoring.

Results

The blinded monitoring procedure is shown to terminate almost surely and to have bounded first and second moments. Its expected sample size is bounded above by the required sample size plus an additional term related to the squared mean difference, representing the cost of blinding. Asymptotically, different behaviors arise: when sigma to infinity, the blinded final sample size approaches the target sample size, whereas when v to infinity, it approaches an inflated limit. In contrast, the unblinded procedure remains asymptotically stable across both regimes. Simulations supported these findings and showed greater overshooting and variability under blinded monitoring in some settings.

Discussion

Blinded continuous monitoring is theoretically justified and practically useful, but it is not always equivalent to unblinded monitoring. Blinding can introduce a measurable sample size penalty and may lead to conservative overshooting under certain asymptotic conditions.

Forward to the past. A frequentist perspective on the Bayesian Time Machine

16:15-16:30

Presenter/Author: Stephen John Senn

Affiliations: University of Sheffield, Sheffield; Medical University of Vienna, Vienna; University of St Andrews, St Andrews

Introduction

One of the claimed advantages of flexible designs is the ability to use data from subjects allocated to act as controls for 'early' treatment arms to augment the data from patients recruited to act as controls for 'late' treatment arms. This, it is claimed, will result in greater efficiency. However, a problem that has been recognised in such trials is that naive contrasts, for example, involving simple pooling of all control patients, can result in biased estimates if there is a time trend. It has been suggested that a Bayesian Time Machine can deal with such trends, thus permitting efficient estimation at a risk of at most minimal bias.

Methods

However, a standard approach to blinding pharmaceutical industry trials is to use placebos that are specific to the active treatment being trialed. If two active treatments are to be compared, the double dummy approach of using a placebo to each is standard. It is thus highly likely that the treatments later in a trial will be recogniseably different in appearance to those used earlier. The flexible trial as a whole is then what has been referred to as veiled: patients do not know what treatment they are receiving but do know that they are not receiving at least one treatment that is in the trial. In fact, eras of changed recruitment are analogous to blocks in incomplete block designs and since randomisation does not occur between eras, such 'blocks' must be regarded as fixed if trend effects are to be eliminated and blinding claimed.

Results

The points are illustrated by calculating frequentist variance inflation factors and potential biases from a published numerical example illustrating the Bayesian Time Machine.

Discussion

It is concluded, that if concurrent control is understood as the key to blinding trials and dealing with trend effects, the frequentist approach of declaring a block whenever the allocation rule is changed and then fitting blocks as fixed effects is the safest for eliminating bias, albeit at the loss of some efficiency, and alternatives should be very carefully and cautiously evaluated.

Anchoring the Winner: A Refined Conditional Test for Enhancing Efficiency in Multi-Arm Selection Trials

16:30-16:45

Presenter: Marco Ratta

Authors: Marco Ratta [1]; Pavel Mozgunov [2]

Affiliations: [1] Saryga, Tournus; [2] MRC Biostatistics Unit, University of Cambridge, Cambridge

Introduction

Multi-arm multi-stage (MAMS) trials with treatment selection are increasingly used to accelerate drug development by comparing multiple experimental treatments against a shared control within an adaptive framework. In "pick-the-winner" designs, one treatment is selected at interim to continue, but this creates multiplicity and selection bias that require strict control of the family-wise error rate (FWER). Existing methods, such as the Stallard and Todd (2003) framework (ST), provide valid inference but can be conservative because they rely on the marginal distribution of the interim winner rather than the realized competitive context.

Methods

The paper considers a two-stage randomized multi-arm trial with multiple experimental treatments and a shared control. At the interim analysis, the arm with the highest interim statistic is selected. The proposed Rank-Based (RB) method conditions inference on the observed runner-up arm, which serves as a "stochastic floor." A conditional tail probability is then calculated for the selected arm, reflecting both its interim ranking and final cumulative evidence, and it is ultimately used to construct a formal hypothesis tesing. Strong control of the FWER is theoreticaslly proven under the global and partial null scenarios, and Monte Carlo simulations comparing RB and ST across several configurations (varying number of arms, information fraction and treatment effects) are performed.

Results

The RB procedure achieves exact FWER control theoretically and empirically, with rejection rates close to the nominal level under the global null setting. When only one treatment is active, RB and ST show similar power. In more competitive scenarios, i.e. when more than one arm exhibits sign of activity, RB provides substantial gains. Power improvements range from about 10% in 3-arm settings with late interim analysis to 15% in 5-arm settings with early interim analysis.

Discussion

By conditioning on the realized runner-up, the RB method uses interim competition as inferential information rather than treating it solely as a feature of the selection process. This improves power and efficiency without sacrificing Type I error control, making the approach especially attractive for high-competition adaptive trials.

A Practical Framework for Using Flexible Parametric Models in the Primary Analysis of Complex Adaptive Trials with Time-to-Event Outcomes: With Applications to Renal Cancer

16:45-17:00

Presenter: Babak Choodari-Oskooei

Authors: Babak Choodari-Oskooei; Elena Frangou; Angela Meade; Mahesh Kb Parmar

Affiliations: UCL Innovative Clinical Trial Unit, London

Introduction

Time-to-event clinical trials are most commonly designed and analysed using the hazard ratio (HR) estimated under a proportional hazards (PH) assumption via a pre-specified Cox model. Flexible parametric models (FPMs) provide a modern alternative, using restricted cubic splines to model baseline hazards and allowing more flexible estimation of treatment effects. Despite their advantages, practical guidance for applying FPMs as the primary analysis method - especially in complex settings such as multi-arm multi-stage (MAMS) trials - remains limited. We introduce a novel structured framework for implementing FPMs, focusing on testing for time-varying effects and selecting spline specifications for modelling the control-arm baseline hazard.

Motivating Example: We illustrate the framework using RAMPART, a MAMS phase III platform trial of immunotherapy in renal cancer. Participants were randomised to control and two research arms, with disease-free survival as the primary outcome. The familywise Type I error rate was strongly controlled at a one-sided 0.025 across both comparisons.

Methods

We provide a step-by-step guide to using FPMs across the trial lifecycle. During the design stage, we show how FPMs fitted to RAMPART control-arm informed power and timelines calculations. At the analysis stage, we describe how to select the FPM, assess the PH assumption using graphical checks and likelihood ratio tests, and choose between reporting HRs (under PH) or restricted mean survival time (under non-PH). We compare the FPM and Cox model fit. Our proposed workflow integrates these steps in a way that ensures control of the design's operating characteristics.

Results

FPMs provide an excellent fit to the survival distribution in all arms. Estimated treatment effects and P-values are consistent with those from the Cox model and logrank tests. Unlike the Cox model, which evaluates the PH assumption separately, FPMs use a single coherent model to generate treatment effects, test statistics, and survival curves.

Discussion

This framework offers practical guidance for adopting FPMs as the primary analysis method in time-to-event trials. By using a unified modelling strategy, it enables clear estimands, coherent summaries, and flexible assessment of PH assumptions, shifting emphasis from empirical Kaplan-Meier curves to model-based survival predictions aligned with trial objectives.

Infectious Disease Modelling

Mon, 28.09.26, 15:30–17:00, Room: N1

Chair: Pio Blieske

EpiParamR: A Unified R Framework for MultiMethod Parameter Estimation in Epidemic Models Using Frequentist, Bayesian, Filtering, and Machine Learning Approaches

15:30-15:45

Presenter: Hom Nath Dhungana

Authors: Hom Nath Dhungana [1, 2]; Sudha Yadav [3]

Affiliations: [1] Charles Sturt University, Australia & Southern Cross Institute, Sydney, Australia, Sydney; [2] Think Data Solution, Kathmandu; [3] Gorakhpur University, Gorakhpur

Introduction

Accurate parameter estimation is central to epidemic modelling, influencing predictions of disease spread and the effectiveness of intervention strategies. Traditional approaches often rely on a single estimation paradigm, limiting robustness and comparative insights across methods. With the increasing availability of epidemiological data, there is a growing need for flexible tools that can integrate multiple estimation techniques within a unified framework. To address this gap, we develop EpiParamR, an R package that enables systematic parameter estimation using a wide range of statistical and machine learning approaches. The package is designed to support reproducible research and facilitate method comparison under diverse epidemic scenarios.

Methods

EpiParamR integrates over twenty estimation techniques grouped into four categories: (i) frequentist methods, including least squares, maximum likelihood, and profile likelihood (ii) Bayesian methods, including Markov Chain Monte Carlo (MCMC), Approximate Bayesian Computation (ABC), Integrated Nested Laplace Approximation (INLA), and Variational Bayes (iii) filtering and state-space approaches, such as Kalman filters, extended and ensemble Kalman filters, and particle filters and (iv) optimisation and machine learning techniques, including simulated annealing, genetic algorithms, neural differential equations, and surrogate-based optimisation. Users provide epidemic time-series data, and the package applies selected methods to estimate model parameters (e.g., recovery rates). Standardised outputs include parameter estimates, uncertainty intervals, model fit statistics, and computational performance metrics.

Results

Simulation experiments and real-world datasets demonstrate that estimation methods produce varying parameter values and predictive accuracy depending on data quality and model structure. Bayesian and filtering approaches generally provide better uncertainty quantification, while optimisation and

machine learning methods show improved performance in complex, nonlinear settings. Frequentist methods offer computational efficiency but may be sensitive to noise and initial conditions. The package enables direct comparison across methods, highlighting trade-offs between accuracy, uncertainty, and computational cost.

Discussion

EpiParamR provides a unified platform that bridges methodological silos in epidemic modelling. By allowing researchers to compare multiple estimation techniques within a consistent framework, it enhances transparency and robustness in parameter inference. The results suggest that no single method is universally optimal instead, method selection should depend on data characteristics and modelling objectives. Future extensions include hybrid estimation strategies and integration with real-time surveillance systems.

Varying Levels of Susceptibility in Epidemic Models

15:45-16:00

Presenter: Tianjiao Yan

Authors: Tianjiao Yan [1]; Hein Putter [2, 1]; Marta Fiocco [1, 3, 2]; Jacco Wallinga [4, 2]

Affiliations: [1] Leiden University, Leiden; [2] Leiden University Medical Center, Leiden; [3] Princess Máxima Center for Pediatric Oncology, Utrecht; [4] National Institute for Public Health and the Environment, Bilthoven

Introduction

Standard epidemiological models typically assume susceptibility is homogeneous however, in real-world populations the risk of infection varies among individuals. Accounting for host heterogeneity can alter the course of epidemics. We investigate infection dynamics inpopulations stratified by levels of susceptibility, in which infectivity acts differently across subgroups. Consequently, highly susceptible individuals leave the susceptible population faster than those with lower susceptibility. This selection mechanism modifies transmission dynamics at the population level, even when individual-level parameters remain constant over time.

Methods

Within the Susceptible-Infected-Removed (SIR) framework, we derive a low-dimensional aggregated epidemic model that incorporates variability in susceptibility, while reproducing the population-level trajectories of the structured model. Derivations show that changes in the composition of susceptibility within the remaining susceptible population lead to time-varying 'apparent' transmission rates and infection risks. We then decompose the effect of heterogeneity into two pathways: the direct pathway which corresponds to the reduction in average frailty induced by selection, and the indirect pathway which corresponds to the lower proportion of the population becoming infected compared with a homogeneous scenario. We further extend the aggregated model to incorporate

time-varying interventions that modify transmission and examine the resulting dynamic changes in risk ratios.

Results

The aggregation model provides a concise and interpretable characterisation of susceptibility heterogeneity. The selection process generates time-varying apparent transmission and infection risk at the population level. The proposed counterfactual decomposition distinguishes between: (i) direct effects, driven by depletion of individuals with higher susceptibility levels, which modulate instantaneous infection risk (ii) indirect effects, driven by feedback loops affecting the rate of new infections through changes in the proportion of infections. Under time-varying interventions, even when intervention effects act multiplicatively, heterogeneity in susceptibility causes risk ratios to evolve over time as the composition of the remaining susceptible population shifts across different intervention scenarios.

Discussion

Our aggregated model and direct/indirect decomposition give insights into how heterogeneity in susceptibility influences the spread of infection within a population.

Integrating Infectious Disease Modelling in a Multi-State Framework

16:00-16:15

Presenter: Liesbeth de Wreede

Authors: Yuwen Ding; Hein Putter; Liesbeth de Wreede

Affiliations: Leiden University Medical Center, Leiden

Introduction

Compartmental epidemic models, notably the susceptible-infected-recovered (SIR) model, are central to describing and forecasting infectious disease dynamics. These models provide mechanistic insight into transmission and intervention effects but typically rely on strong parametric assumptions. In contrast, multi-state models as used in survival analysis to study processes allow non-parametric estimation of transition hazards and transition probabilities and easy extension of the state space. We hypothesized that integrating these two approaches would yield flexible and reliable models for infection-related processes and that this would in particular be useful for assessing the impact of intervention strategies, e.g., earlier vaccination.

Methods

Our starting point was a non-parametric Markov multi-state model in which hazards were estimated by means of the Nelson-Aalen estimator and transition probabilities by the Aalen-Johansen (AJ)

estimator. We developed two generalizations of the SIR model, a mechanistic ordinary differential equations (ODE) model assuming (piecewise) constant parameters in which estimation is MLE-based, and a hybrid approach that retains the non-parametric structure while replacing AJ-estimates of infection-related hazards by MLE-based estimates. The performance of the three methods was compared in an extensive simulation study, based on an extended stochastic SIR model with vaccination and death compartments. Two baseline settings (constant and time-varying vaccination rate) and three intervention scenarios are considered: doubling the vaccination rate, scaling infection rates by 0.75 to represent lockdown, and scaling infection rate after vaccination by 0.75 to reflect improved vaccine protection.

The different models are also fitted on population-based data from the COVID-19 pandemic in the Netherlands.

Results

In well-specified models, all approaches perform similarly in the baseline settings, with a variance advantage for the ODE model. In the intervention scenarios, the AJ estimator, lacking an explicit transmission mechanism, fails to extrapolate to unobserved settings. The hybrid and MLE-based approach are able to predict counterfactual changes accurately, but the latter performs less well with time-varying hazards due to its constant hazard-assumption for the non-infection-related transitions.

Discussion

Our study demonstrates the potential of a hybrid strategy bridging multi-state and compartmental epidemic modelling which correctly takes into account the dependence between individuals in an infectious disease setting.

"Correcting Length-Sampling Bias and Interval Censoring in Point Prevalence Surveys Using Multi-State Models"

16:15-16:30

Presenter/Author: Maria Eugenia Messuti

Affiliations: IMBI, Freiburg, freiburg

Introduction

Hospital-acquired infections (HAIs) represent a major global public health challenge [1]. While prospective cohort studies remain the gold standard, Point Prevalence Surveys (PPS) are widely used due to their cost-effectiveness [2]. Extended PPS protocols employed by agencies such as the CDC and ECDC introduce complex methodological challenges, including immortal time bias, competing risks, interval-censored infection times, and length-sampling bias, preventing direct application of standard survival methods [3]. This work develops and validates the first integrated statistical framework addressing all these challenges simultaneously.

Methods

Using an anonymized real-world ICU cohort of 2,126 patients [4], PPS datasets were mimicked by applying protocol-specific sampling procedures to the existing longitudinal data. A Simple Illness-Death Multi-state model served as the analytical framework, with a Full Cohort Analysis (FCA) establishing unbiased reference estimands. Two PPS protocols were mimicked 100 times: a standard PPS and a CDC-style protocol introducing additional interval censoring. IPW based on the inverse length of stay corrected for length-sampling bias [5], while interval censoring was handled via a parametric Multi-State Model using the msm R package [6]. Performance was evaluated using Cumulative Hazard Functions (CHFs), Stacked Probability Plots (SPPs), and covariate Hazard Ratios (HRs).

Results

Unweighted PPS estimates consistently deviated from FCA reference values. IPW correction successfully realigned CHFs, SPPs, and HRs with the unbiased reference for both protocols. For the CDC protocol, IPW-weighted estimates centred on the true estimands, though with wider uncertainty reflecting interval censoring. A single-run CDC analysis confirmed practical applicability, with SPPs remaining robust after correction.

Discussion

This work establishes the first comprehensive multi-state modelling strategy correcting for all major sources of bias in extended PPS data simultaneously. The IPW-MSM framework reliably recovers unbiased estimates on average, with practical R code provided for implementation. Future work should extend the methodology to EPIC, SPREAD, and ECDC protocols.

References

[1] WHO. Infection prevention and control guidelines. 2016.

[2] Vincent JL et al. JAMA, 302:2323-2329, 2009.

[3] Schumacher M et al. Int J Epidemiol, 42:1502-1508, 2013.

[4] Savin I et al. J Crit Care, 45:95-104, 2018.

[5] Staus P et al. Clin Epidemiol, 14:1053-1064, 2022.

[6] Jackson C. J Stat Softw, 38:1-28, 2011.

A smooth hierarchical model for parasite and gametocyte dynamics in controlled human malaria infection reveals population-dependent parasite-gametocyte coupling

16:30-16:45

Presenter: Jordache Ramjith

Authors: Jordache Ramjith [1]; Luc Coffeng [2]; Matthew McCall [1]; Issaka Sagara [3]; Teun Bousema [1]

Affiliations: [1] Radboud University Medical Center, Nijmegen; [2] Department of Public Health, Erasmus MC University Medical Center, Rotterdam; [3] Université des Sciences, des Techniques et des Technologies de Bamako, Bamako

Introduction

Controlled human malaria infection (CHMI) studies provide data on early parasite dynamics and gametocyte (transmission stage) emergence, but existing within-host models typically rely on discrete parasite-generation structures that may be difficult to identify from observed data. This complicates inference, particularly for early infection dynamics and for quantifying how parasite kinetics translate into transmission potential. For Plasmodium falciparum, gametocyte production varies substantially between infections and is difficult to analyse because maturation occurs largely outside the bloodstream and is therefore unobserved. We propose a smooth hierarchical modelling framework that avoids explicit generation assumptions while directly linking parasite dynamics to gametocyte production.

Methods

We developed a joint Bayesian hierarchical model for asexual parasite and gametocyte dynamics using continuous-time exponential growth modulated by smooth onset functions. Gametocyte density is represented as a delayed transformation of parasite density. The model includes a regression structure linking gametocyte amplitude to parasite initial density, growth rate, and timing, enabling estimation of multiplicative effects on the natural scale. Between-host variability is captured through hierarchical random effects with weakly informative priors. Observations are modelled using censored log-normal likelihoods to account for stage-specific detection limits, and inference is performed using Hamiltonian Monte Carlo.

Results

The model was applied to CHMI data from malaria-naïve individuals (Netherlands) and to CHMI data from an endemic population (Mali). Stable inference was achieved without requiring assumptions on parasite generation number. Parasite growth rates were higher and onset occurred earlier in Netherlands compared to Mali, where growth was slower and more heterogeneous. In Netherlands, gametocyte production was strongly associated with parasite growth rate and initial density, with approximately 2- to 4-fold increases in predicted gametocyte density across their observed ranges. In contrast, in Mali this dependence was weaker and more variable, with a stronger influence of cumulative parasite exposure and timing, consistent with population-level differences in parasite dynamics.

Discussion

This framework enables direct investigation of how parasite dynamics shape gametocyte production. The results suggest that the relationship between parasite kinetics and transmission potential differs between malaria-naïve and endemic settings, highlighting the importance of accounting for epidemiological context and host immunity when modelling malaria transmission.

A Hierarchical Spatio-Temporal Bayesian Framework for Modelling Dengue Transmission Dynamics in South Asia

16:45-17:00

Presenter: Hom Nath Dhungana

Authors: Hom Nath Dhungana [1, 2]; Masood Siddiqui [3]; Sraddha Sharma [3]

Affiliations: [1] Charles Sturt University, Sydney; [2] Southern Cross Institute, Sydney, Sydney; [3] Lucknow University, Lucknow

Introduction

Dengue fever represents a rapidly expanding arboviral threat across South Asia, characterised by pronounced spatial clustering, seasonal epidemics, and escalating clinical burden. Conventional surveillance analyses often fail to account for spatial dependence, temporal autocorrelation, and environmental forcing. We propose a hierarchical spatio-temporal Bayesian modelling framework to generate clinically actionable risk estimates and improve outbreak prediction.

Methods

District-level dengue case counts from multiple South Asian settings were analysed using a Bayesian hierarchical model with a negative binomial likelihood to address over-dispersion. Spatial dependence was modelled via conditional autoregressive (CAR) priors, while temporal trends were captured using second-order random walk processes. Climate covariates (temperature, rainfall, humidity) were incorporated with distributed lag structures. A spatio-temporal interaction term quantified localised epidemic deviations. Bayesian inference was performed using Integrated Nested Laplace Approximation (INLA). Model performance was assessed via posterior predictive checks, WAIC, and out-of-sample forecasting accuracy.

Results

The model identified persistent high-risk transmission clusters concentrated in densely urbanised districts, alongside emerging peri-urban hotspots. Significant lagged associations were observed between monsoon rainfall, temperature increases, and dengue incidence. Inclusion of spatio-temporal interaction effects improved predictive fit and enabled detection of district-specific epidemic amplification. Forecast validation demonstrated superior accuracy compared with non-spatial and classical time-series models. Scenario analyses indicated that geographically targeted vector control could substantially reduce projected case burdens.

Discussion

This study demonstrates the utility of spatio-temporal Bayesian disease mapping for capturing complex dengue transmission dynamics in climate-sensitive regions. The framework provides robust uncertainty quantification and operationally relevant forecasts to inform clinical surge planning, surveillance optimisation, and precision public-health interventions. Its scalability supports adaptation to other vector-borne diseases and real-time early warning systems across endemic settings.

Digital Health Education and Transformation

Mon, 28.09.26, 15:30–17:00, Room: N2

Chair: Sabine Hoffmann

Observing the Long Struggle for Digital Transformation in German Hospitals: Results from a Panel Study 2017 to 2025

15:30-15:45

Presenter: Florian Kücking

Authors: Florian Kücking; Annika Hering; Jana Strate; Laura Naumann; Jan-David Liebe; Ursula Hertha Hübner

Affiliations: Hochschule Osnabrück - University of Applied Sciences, Osnabrück

Introduction

Despite major policy initiative, Germany's hospital digital transformation remains slow.

Methods

This study analysed data from the IT-Report Healthcare (2017, 2023 and 2025) to evaluate changes in five domains related to the Hospital Future Act (KHZG): clinical documentation, order entry, clinical decision support (CDSS), medication management and discharge management. A panel of 49 hospitals was examined using linear mixed models (LMM).

Results

While most hospitals applied for funding in documentation and discharge management, by 2025, most projects also in these areas were still in progress. Significant improvements to 2017 were observed in medication management ($p = 0.009$ in 2023 $p = 0.044$ in 2025) and discharge management ($p < 0.001$ in 2025), while progress in other areas was limited.

Discussion

The findings indicate that digital maturity increased only partially due to administrative delays and implementation bottlenecks and that a sustained evaluation beyond the KHZG period, including consideration of the implementation context, is needed to capture the long-term effects on clinical digitalisation.

The full paper of this abstract will be published in the GMDS series "Studies in Health Technology and Informatics" with doi: 10.3233/SHTI260989, URL:https://dx.doi.org/10.3233/SHTI260989

Cross-Site Transfer and Adaptation of an In-House CDSS: First Lessons Learned

15:45-16:00

Presenter: Katja Hoffmann

Authors: Katja Hoffmann [1]; Katharina Schuler [1]; Clemens Malkawi [2]; Alexander Twrdik [2]; Martin Federbusch [2, 3]; Toralf Kirsten [4, 2]; Martin Sedlmayr [1]

Affiliations: [1] Institute for Medical Informatics and Biometry, Faculty of Medicine Carl Gustav Carus, TUD Dresden University of Technology, Dresden; [2] Department for Clinical AI and Translational Medicine, University of Leipzig Medical Center, Leipzig; [3] Institute of Laboratory Medicine, Clinical Chemistry and Molecular Diagnostics, University Leipzig Medical Center, Leipzig; [4] Institut für Medizinische Informatik, Statistik und Epidemiologie, Universität Leipzig, Leipzig

Introduction

This study investigates the transfer of core components of an in-house clinical decision support system (CDSS) between two university hospitals in Germany, aiming to derive transferable lessons for CDSS transfer.

Methods

We submit a Technical Case Report and it's structured by ISCIL instead of IMRaD.

Results

We submit a Technical Case Report and it's structured by ISCIL instead of IMRaD.

Discussion

We submit a Technical Case Report and it's structured by ISCIL instead of IMRaD.

State of the Art

While frameworks for CDSS implementation are well established, guidance on transferring in-house systems to independent clinical sites remains limited, particularly from the perspective of the adopting institution.

Concept

The study frames transfer as a structured process of recontextualization, differentiating site-agnostic core components from locally adapted elements and applying the NASSS framework to systematically identify transferable challenges and scalable strategies.

Implementation

Core components were conceptually integrated with local infrastructure via a modular design, including a data ingestion layer and a UI integration layer, enabling site-specific adaptation and seamless EHR embedding. The system is intended to support clinicians managing patients at risk of acute kidney injury.

Lessons learned

Key lessons highlight the importance of modular, adaptable components, transparent data pipelines, tailored user interfaces, early stakeholder engagement, robust regulatory and quality management capacity, and consideration of broader system requirements, including MDR compliance.

The full paper of this abstract will be published in the GMDS series "Studies in Health Technology and Informatics" with doi: 10.3233/SHTI260991, URL:https://dx.doi.org/10.3233/SHTI260991

Inter- and Transdisciplinarity Competencies in Medical Informatics and Digital Health Education - A Scoping Review

16:00-16:15

Presenter: Ekaterina Kutafina

Authors: Ekaterina Kutafina [1]; Jens Dierkes [2]; Oya Beyan [3]

Affiliations: [1] Institute for Biomedical Informatics, Köln; [2] Department of Research and Publication Support, University and City Library, University of Cologne, Cologne; [3] Universität zu Köln, Köln

Introduction

Medical informatics and digital health are intrinsically transdisciplinary fields, requiring practitioners to integrate clinical knowledge, computational methods, regulatory understanding, and societal engagement. While inter- and transdisciplinary (ITD) competencies are widely considered essential in these domains, it remains unclear how explicitly they are addressed in existing training programs.

Methods

To map the current landscape of interdisciplinary and transdisciplinary training initiatives relevant to medical informatics and digital health and to identify gaps and strategies in ITD competency development, a scoping review was conducted following PRISMA-ScR guidelines. PubMed was searched using a query combining terms for inter-/transdisciplinarity, education/training, and computational/informatics domains.

Results

After two-step title and abstract screening, 20 publications were included for full-text analysis. Publications were assessed for disciplinary coverage, ITD competency approaches, pedagogical frameworks, and relevance to ITD training in medical informatics. The 20 included papers show that ITD competencies are predominantly addressed implicitly through project-based or collaborative learning rather than as explicit, assessable learning objectives.

Discussion

A clear gap exists between the acknowledged importance of ITD competencies and their formal integration into medical informatics and digital health curricula. Future work should develop explicit ITD learning objectives, structured curricula, and dedicated assessment methods, aligned with ITD meta-research and supported by institutional and network-level initiatives.

The full paper of this abstract will be published in the GMDS series "Studies in Health Technology and Informatics" with doi: 10.3233/SHTI261018, URL:https://dx.doi.org/10.3233/SHTI261018

Assessing Biases and Discrimination in German Medical Informatics

16:15-16:30

Presenter: Michael Rusongoza Muzoora, Christelle Kemda Ngueda

Authors: Michael Rusongoza Muzoora [1]; Christelle Kemda Ngueda [2]; Johannes Oehm [3]; Tobias Brix [4]; Christina Birgit Behrend [1]; Alessandra Simone Kuntz [5]; Marco Schaarschmidt [1]; Maija Poikela [1]; Sylvia Thun [1]; Carina Nina Vorisek [1]

Affiliations: [1] Berlin Institute of Health at Charité Universitätsmedizin Berlin, Berlin; [2] Institute of Medical Statistics, Computer and Data Sciences. Jena University Hospital., Jena; [3] Universität Münster, Münster; [4] Institute of Medical Informatics, University of Münster, Münster; [5] Universitätsmedizin Göttingen, Medizinische Informatik, Göttingen

Introduction

Biases and discrimination in the professional field limit workplace effectiveness by creating environments that stifle productivity and diminish innovation. While recent research on discrimination in public institutions in Germany has gained public traction, empirical evidence within the medical informatics in Germany remains limited. This study aims to assess the status of perceived biases, discrimination, diversity, and fairness in the field of medical informatics in Germany.

Methods

A 33-item anonymous cross-sectional online survey was conducted by the Medical Informatics, Biometrics, and Epidemiology (GMDS) e.V. (Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (GMDS) e.V.) - Diversity Project Group, targeting individuals working in medical informatics and related fields.

Results

Data was collected using REDCap and included demographic information, self-identification with minority groups, perceptions of biases and discrimination and fairness through Likert-scale items.

Discussion

The study provides a first systematic overview of perceived biases and discrimination by medical informaticians in Germany, establishing a foundation for evidence-based strategies for professional societies and research institutions.

The full paper of this abstract will be published in the GMDS series "Studies in Health Technology and Informatics" with doi: 10.3233/SHTI261017, URL:https://dx.doi.org/10.3233/SHTI261017

Identification of Nursing Informatics Competencies for Ambulatory Care Nurses in Germany: A Cross-Sectional Survey

16:30-16:45

Presenter: Nicole Egbert

Authors: Annika Hering; Nicole Egbert; Florian Kücking; Ursula Hertha Hübner

Affiliations: Hochschule Osnabrück, Osnabrück

Introduction

Ambulatory nursing care is increasingly shaped by digital transformation, which requires context-specific competencies. This study aimed to identify the relevance of nursing informatics competencies in the ambulatory care setting in Germany.

Methods

A cross-sectional, anonymous online survey was conducted in which 43 competency areas were rated across five nursing roles using a 0 to 100% scale. Descriptive analyses were performed to identify the top five core competency areas for each role.

Results

Across all roles, data protection and privacy competency was consistently rated as highly relevant.

Discussion

The findings emphasize the need for targeted education and support the development of a context-specific competency framework.

The full paper of this abstract will be published in the GMDS series "Studies in Health Technology and Informatics" with doi: 10.3233/SHTI261016, URL:https://dx.doi.org/10.3233/SHTI261016

Addressing Challenges in Asynchronous Online Biostatistics Teaching: Redesign of a Master of Public Health Module

16:45-17:00

Presenter: Lára R. Hallsson

Authors: Lára R. Hallsson [1]; Sibylle Puntscher [1]; Marjan Arvandi [1]; Uwe Siebert [1, 2, 3]; Gaby Sroczynski [1]

Affiliations: [1] Department of Public Health, Health Services Research and Health Technology Assessment, UMIT TIROL - University for Health Sciences and Technology, Hall in Tirol; [2] Center for Health Decision Science, Departments of Epidemiology and Health Policy & Management, Harvard T. H. Chan School of Public Health, Boston; [3] Institute for Technology Assessment and Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston

Introduction

Teaching biostatistics in an asynchronous online setting poses considerable didactic challenges. The module Biostatistics I is part of the first semester in the Master Public Health program (4 semesters, 120 ECTS) at UMIT TIROL. It was designed in 2019 as a three-week asynchronous online module using the learning platform Moodle. Delivery of the module revealed challenges such as limited student interaction, low forum participation, hesitation to ask questions, and limited insight into students' learning progress. The use of statistical software without immediate feedback hindered learning. Thus, the aim was to redesign the module to address challenges related to asynchronous teaching of biostatistics.

Methods

Based on student evaluations and feedback, the module was re-designed iteratively over the past five years. While retaining the online format, the module duration was extended from three to six weeks and structured into three two-week blocks combining asynchronous learning with optional synchronous elements. A face-to-face kickoff session introduced course objectives, structure, deadlines, and assessment procedures as well as the statistical software. During the online phase, students worked with lecture videos, forum assignments, and statistical software exercises that integrated theory with practical application. Tasks were completed individually or in groups and shared in Moodle forums for peer and instructor feedback. Weekly optional synchronous tutorials via Zoom were introduced.

Results

Overall the success of the module depended strongly on students' acceptance of the online format. The face-to-face kickoff session fostered interaction, reduced anonymity, encouraged students to ask questions, and lowered barriers working with statistical software, resulting in increased students' interactions throughout the module. Participation and engagement in the optional Zoom tutorials were high. Students used opportunities for discussion, exercises, and clarification of key topics. Expanding synchronous tutorials, particularly through weekly alternating sessions focused on lecture content and statistical software practice, further enhanced learning support and students' satisfaction.

Discussion

Asynchronous online teaching of biostatistics can be effective when supported by clear structure, early interaction, and regular feedback. Combining guided self-study with synchronous tutorials and an initial face-to-face session helped increase engagement and reduce learning barriers. Continuous adaptation based on student feedback remains essential to further improve learning outcomes.

Network Meta-Analysis and Evidence Reliability

Mon, 28.09.26, 15:30–17:00, Room: N4

Chair: Virginia Chiocchia

Assessment of inconsistency and its association with network structural characteristics in 177 network meta-analyses of continuous data

15:30-15:45

Presenter: Myanca Rodrigues

Authors: Myanca Rodrigues [1]; Sofia Tsokani [2]; Eirini Pagkalidou [2]; Balpreet Panesar [3]; Menelaos Konstantinidis [4]; Andrea C. Tricco [5]; Areti-Angeliki Veroniki [6]

Affiliations: [1] Queen's University, Kingston; [2] Department of Hygiene, Social-Preventive Medicine and Medical Statistics, School of Medicine, Aristotle University of Thessaloniki, Thessaloniki; [3] Toronto Metropolitan University, Toronto; [4] Institute of Health Policy, Management & Evaluation, Dalla Lana School of Public Health, University of Toronto, Toronto; [5] St. Michael's Hospital, Unity Health Network, Toronto; [6] Brown University, Providence

Introduction

Network meta-analysis (NMA) allows the comparison of multiple interventions within a single model but relies on the assumption of consistency between direct and indirect evidence. Empirical evaluations of inconsistency have focused primarily on NMAs with binary outcomes. Continuous outcomes represent approximately one-third of published NMAs, yet their structural and methodological characteristics and the prevalence of inconsistency remain understudied. We aimed (i) to describe the characteristics of NMAs with continuous primary outcomes and (ii) to assess the prevalence of global inconsistency in continuous network NMAs.

Methods

We used a previous compiled database of published NMAs of randomized controlled trials (searches through 27 June 2018). We extracted publication, network, and model characteristics and summarized them descriptively. In networks with at least one closed loop and available outcome data, global inconsistency was assessed using the design-by-treatment interaction model. Analyses were stratified by effect measure (mean difference [MD] and standardized mean difference [SMD]). Evidence against consistency was evaluated at significance levels of 0.05 and 0.10.

Results

From 1144 eligible NMAs, we identified 177 with continuous primary outcomes. Among 177 NMAs, 73% were full networks (≥1 loop) and 15% were open networks. The number of NMAs with continuous outcomes increased from 2013 though 2015. Most evaluated efficacy outcomes (96%). Random-effects models (65%) and Bayesian approaches (72%) were common among full shaped

networks. Reporting of transitivity, heterogeneity exploration, and publication bias assessment was <50% across NMAs.

Among networks with loops, evidence against global consistency at $p<0.05$ was observed in 19/38 (50%) networks with MD effect size and in 16/32 (50%) networks with the SMD effect size. At $p<0.10$, inconsistency was detected in 21/38 (55%) MD networks and 17/32 (53%) SMD networks.

Discussion

NMAs with continuous outcomes have become increasingly common and demonstrate substantial variability in network structure and statistical methodology. Evidence of global inconsistency was observed in approximately half of networks with closed loops, irrespective of effect measure. These findings highlight the importance of routine inconsistency assessment and transparent methodological reporting in NMAs of continuous outcomes.

Assessing the Impact of Model Assumptions in Network Meta-Regression: A Simulation Study

15:45-16:00

This is one of this year's ISCB StCA award-winning abstracts.

Presenter: Nana-adjoa Kwarteng

Authors: Nana-adjoa Kwarteng [1]; Guido Schwarzer [1]; Adriani Nikolakopoulou [2, 1]; Theodoros Evrenoglou [1]

Affiliations: [1] Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center-University of Freiburg, Freiburg; [2] Department of Hygiene, Social-Preventive Medicine and Medical Statistics, School of Medicine, Aristotle University of Thessaloniki, Thessaloniki

Introduction

Network meta-regression (NMR) models extend network meta-analysis (NMA) by simultaneously synthesizing evidence on multiple treatments while adjusting for effect modifiers. By accounting for effect modification, NMR can reduce between-study heterogeneity relative to NMA. However, available modeling assumptions in NMR lead to several models that address related but non-identical research questions, complicating model selection. The consequences of model misspecification under varying network structures, levels of heterogeneity, and interaction assumptions remain unclear.

Methods

To evaluate the performance of alternative NMR models and provide practical recommendations, we conducted a simulation study, varying key network characteristics, including network sparsity, study designs, and heterogeneity magnitude. NMR was examined under frequentist and Bayesian frameworks. We considered the standard NMA model without interactions, as well as four NMR parameterizations differing in across-comparison interaction assumptions (common vs. independent interactions) and interaction consistency assumptions (with or without consistency).

Results

Combining network characteristics with true model specifications resulted in 120 simulation scenarios. In fully connected networks of two-arm studies with low heterogeneity and without effect modification, all NMR models demonstrated low bias and mean squared error. In contrast, NMA models overestimated treatment effects and heterogeneity in the presence of effect modification. NMR models with independent across-comparison interactions generally achieved nominal coverage and low bias when the data-generating mechanism for interactions was consistent and common across comparisons. However, their coverage deteriorated in sparse networks with between-study heterogeneity. In corresponding scenarios, models assuming common across-comparison interactions were more robust to sparsity and heterogeneity but yielded conservative inference for interaction effects. In networks containing multi-arm studies, NMR models lacking interaction consistency showed sub-nominal coverage in their corresponding data-generating mechanisms.

Discussion

Our findings reinforce that ignoring treatment-covariate interactions by relying on standard NMA may yield biased treatment effect and heterogeneity estimates. In the presence of effect modification, NMR models are preferable and require careful consideration of modeling assumptions, anticipated patterns of effect modification, and the structure of the evidence network. Models assuming consistent interactions are advantageous in networks with multi-arm studies. Sparse networks often necessitate common-interaction assumptions. These findings provide guidance for planning and interpreting network meta-analyses and meta-regressions in complex treatment networks.

Sequential treatment strategies in component network meta-analysis: a statistical framework and application

16:00-16:15

Presenter: Adriani Nikolakopoulou

Authors: Adriani Nikolakopoulou; Garo Garabetian; Myrto Samara; Anna-Bettina Haidich

Affiliations: Aristotle University of Thessaloniki, Thessaloniki

Introduction

Network meta-analysis (NMA) enables comparison of multiple interventions but typically treats treatments as static and independent of their order of administration. This assumption may be inadequate in many clinical settings where treatments are delivered sequentially and effects depend on prior exposure. Component network meta-analysis (CNMA) extends NMA by modelling interventions as combinations of components however, existing approaches do not account for treatment sequencing. We aimed to develop a statistical framework to evaluate sequential treatment strategies within CNMA.

Methods

We propose a generalized CNMA framework based on a graph-theoretical representation of treatment networks to model interventions delivered across multiple phases. Treatment pathways are decomposed into phase-specific components, allowing estimation of sequence-dependent effects and non-commutative treatment ordering (e.g., A→B ≠ B→A). Multi-arm trials are incorporated while accounting for correlations between comparisons. The model uses a generalized least squares estimation approach within an extended network meta-analysis framework, and builds on connections between NMA, electrical networks, and flow-based representations of evidence.

Results

The proposed sequential CNMA model was evaluated using simulated networks with multiple treatment phases to assess identifiability and estimability of phase-specific parameters. The framework was then applied to a network of clinical trials comparing two-phase antipsychotic treatment strategies for schizophrenia. The model enabled estimation of phase-specific effects and comparison of alternative treatment pathways, demonstrating its ability to capture temporal dynamics not addressed by standard NMA or CNMA approaches.

Discussion

This work extends CNMA to explicitly incorporate treatment sequencing, addressing an important limitation of current evidence synthesis methods. By allowing estimation of sequence-specific and phase-dependent effects, the framework provides a more realistic representation of clinical pathways. The approach may support decision-making in settings where treatment effects evolve over time, although further work is needed to explore model assumptions, data requirements, and applicability in more complex networks.

A Bayesian framework for continuously updating network meta-analyses

16:15-16:30

Presenter: Katharina Prinz

Authors: Katharina Prinz; Anna Chaimani

Affiliations: Oslo Center for Biostatistics and Epidemiology, Department of Biostatistics, University of Oslo, Oslo

Introduction

Living systematic reviews and meta-analyses are commonly used to provide up-to-date evidence summaries for specific outcomes of interest. However, repeatedly updating meta-analyses results as new evidence becomes available may lead to inflated type I error. Sequential methods for clinical trials have been adapted to pairwise meta-analysis to address this problem. These methods, though, are not directly applicable to network meta-analysis, where multiple correlated treatment effects, indirect evidence, and uncertainty about inconsistency need to be properly considered.

Methods

We extend a Bayesian sequential meta-analysis approach proposed previously for pairwise meta-analysis to the context of random effects network meta-analysis. We use informative prior distributions for heterogeneity and inconsistency based on previous empirical evidence. We then apply stopping rule criteria directly to the posterior distribution for the basic treatment effect parameters. Specifically, at the first interim analysis, posterior distributions are obtained for the vector of relative effects theta, the heterogeneity, and inconsistency. The posterior distribution for theta is then subject to stopping rules the latter consider both direct and indirect evidence for every treatment contrast. At every update, the likelihood is recalculated using the updated available data, and the model is fitted as previously by employing the previous posterior distribution as the next prior.

Results

We illustrate our approach using data from a living network meta-analysis on systemic pharmacologic treatments for chronic plaque psoriasis. We compare the results and conclusions from every update with and without adjustment. We also monitor how many updates are needed to obtain conclusive evidence for every relative effect when using sequential pairwise or network meta-analysis. Results will be presented at the conference.

Discussion

The proposed framework extends the Bayesian sequential meta-analysis approach to network meta-analysis, which allows preserving sequential validity while taking advantage of the network structure by combining direct and indirect evidence.

Learning study similarity to investigate heterogeneity in meta-analysis using LLMs and triplet loss

16:30-16:45

Presenter: Kanella Panagiotopoulou

Authors: Kanella Panagiotopoulou; Theodoros Evrenoglou

Affiliations: Institute of Medical Biometry and Statistics (IMBI), Faculty of Medicine and Medical Center, University of Freiburg, Freiburg im Breisgau

Introduction

Meta-analyses of observational studies are frequently characterized by substantial between-study heterogeneity, which limits the interpretability of pooled estimates. Although conventional meta-regression can, in principle, be used to explore heterogeneity, it is often underpowered to handle multiple effect modifiers. We propose a novel framework that integrates large language models (LLMs) with deep metric learning to infer study-level similarity and uncover latent structural patterns among studies prior to meta-analysis.

Methods

Study-level clinical and methodological characteristics were processed by an LLM to generate relative similarity constraints in the form of study triplets (anchor, similar, dissimilar). These triplets were constructed by treating each study as an anchor and comparing it with pairs of other studies to identify, in each instance, the study most similar to the anchor. We use the resulting study triplets to train a neural network using the triplet loss algorithm, a deep learning approach widely applied in fields such as computer vision to learn similarities between data points. In our context, this algorithm learns an embedding space where clinically and methodologically similar studies are clustered together. We apply our framework to a published meta-analysis dataset of 58 observational studies comparing cognitive outcomes between preterm- and term-born children. Subsequently, we fit meta-analysis models within the identified study clusters and compare the results with those of the overall analysis.

Results

The results revealed clear structure, with two dominant clusters and a small set of studies whose characteristics did not align with any cluster and were identified as noise. One cluster exhibited reduced between-study heterogeneity, with substantially narrower confidence and prediction intervals compared to the overall analysis, whereas the second cluster retained higher variability. Pooled effect estimates differed across clusters, suggesting context-specific effects that were obscured in the overall analysis.

Discussion

This work presents a novel approach for exploring heterogeneity in meta-analysis. By transforming unstructured study characteristics into a similarity space, our framework enables the identification of coherent subgroups. This approach extends beyond current practice, where heterogeneity is typically investigated only after the primary analysis. Overall, the proposed framework supports more precise inference, thereby enhancing decision-making in the presence of heterogeneous real-world evidence.

From Poor Agreement to Context-Dependent Reliability: A Two-Year Evolution of AI in Risk of Bias Assessment for Systematic Reviews

16:45-17:00

Presenter: Tanyeli Guneyligil Kazaz

Authors: Tanyeli Guneyligil Kazaz [1]; Arzu Baygul Eden [2]; Didem Derici [3]; Seval Kul [1, 4]

Affiliations: [1] Gaziantep University, School of Medicine, Department of Biostatistics, Gaziantep; [2] Koç University, School of Medicine, Department of Biostatistics, İstanbul; [3] Mersin University, School of Medicine, Department of Biostatistics and Medical Informatics, Mersin; [4] Cardiovascular Research Institute Basel (CRIB), Basel

Introduction

Risk of bias assessment is a key component of systematic reviews, traditionally requiring expert judgment. Moens et al. (2025) reported poor agreement between AI tools and human reviewers, with near-zero or negative inter-rater reliability, highlighting major limitations of earlier AI systems. This study evaluates how AI performance has evolved over two years and whether recent models can approach expert-level reliability.

Methods

We replicated risk of bias assessments from Moens et al. (2025), based on tools available in March 2024, using updated large language models and multiple evaluation modes (Standard, Balanced, Advanced). Three tools were evaluated: RoB-2, Newcastle-Ottawa Scale and Downs and Black checklist. Agreement with manual (ground truth) assessments was quantified using percentage agreement and risk-level concordance across models, including ChatGPT (v5.3, GPT-4-based) and Gemini.

Results

Performance varied by model and evaluation mode. For Downs and Black, ChatGPT 5.3 and GPT-4-based models showed high agreement (83.3%), while Gemini and Balanced mode were moderate (66.7%). For Newcastle-Ottawa Scale, Advanced mode reached perfect agreement (100%), whereas Standard (47.6%) and Balanced (66.7%) overestimated study quality, particularly in lower-quality studies. For RoB-2, Standard and Balanced modes showed high agreement (91.7%), while Advanced mode had lower overall agreement (33.3%) but better identification of high-risk studies (83.3%). Gemini achieved overall agreement of 80.8%, with strong detection of high-risk studies but inconsistent classification of low-risk studies.

Discussion

AI performance varies across contexts but shows clear improvement over earlier reports. Reliability depends on tool structure: highest for structured Newcastle-Ottawa Scale, intermediate for Downs and Black, and most variable for RoB 2, which requires complex, context-dependent judgment. Stricter evaluation increases sensitivity but reduces specificity, leading to overestimation of bias. Overall, AI performs well on structured tools but remains limited for complex frameworks. While modern models can approach human agreement under optimized conditions, they are not yet suitable as standalone tools. Future work should focus on calibration, evaluation-mode optimization, and domain-specific adaptation to improve robustness and consistency.

Reference:

Moens M, Nagels G, Wake N, Goudman L. (2025). Artificial intelligence as team member versus manual screening to conduct systematic reviews in medical sciences. iScience.

Dynamic Survival Modeling and Prediction

Mon, 28.09.26, 15:30–17:00, Room: Runder Saal

Chair: Giuliana Cortese

Cluster-weighted modeling of time-to-event hierarchical data for profiling COVID-19 heart failure patients

15:30-15:45

Presenter: Francesca Ieva

Authors: Luca Caldera [1]; Andrea Cappozzo [2]; Chiara Masci [3]; Anna Paganoni [1]; Francesca Ieva [4, 5]

Affiliations: [1] MOX - Modelling and Scientific Computing Lab, Department of Mathematics, Politecnico di Milano, Milano; [2] Università Cattolica del Sacro Cuore, Milano; [3] Università degli Studi di Milano, Milano; [4] MOX - Modelling and Scientific Computing Lab, Department of Mathematics, Politecnico di Milano, Milan; [5] Health Data Science Research Centre, Human Technopole, Milan, Italy, Milano

Introduction

The COVID-19 pandemic has placed extraordinary pressure on healthcare systems, particularly affecting patients with heart failure (HF), who show increased vulnerability and mortality when infected with SARS-CoV-2. In severely affected regions such as Lombardy (Italy), understanding survival requires models that account for both patient-level heterogeneity and hospital-level variability. Standard survival models typically address only one of these aspects, rarely integrating both within a unified clustering framework.

Methods

We propose a novel multilevel Cluster-Weighted Model (CWM) for right-censored time-to-event data, combining mixture modeling with parametric shared frailty survival analysis. CWMs flexibly model covariates and responses in heterogeneous populations, but existing extensions do not handle survival outcomes. Our approach incorporates a parametric frailty term within each latent component, enabling simultaneous identification of latent patient profiles and cluster-specific survival mechanisms while accounting for hospital-level heterogeneity. Within each cluster, survival follows a parametric shared frailty model including individual covariates and a hospital-level random effect. Alternative parametric baseline hazards are considered. Parameters are estimated via two EM-based algorithms - Classification EM (CEM) and Stochastic EM (SEM) - adapted for right-censored data. Model selection relies on the Bayesian Information Criterion.

Results

The method is applied to data on 3,086 HF patients hospitalized with COVID-19 in Lombardy (January 2020-June 2021) across 32 hospitals, with a 90-day follow-up. Covariates (age, gender, comorbidity score, COPD, bronchitis) drive clustering, while pneumonia and respiratory failure enter the survival component. Three clusters are identified, with a Lognormal baseline hazard. The first cluster includes highly comorbid patients with the poorest survival and minimal hospital variability. The second cluster is characterized by respiratory disease, with pneumonia increasing mortality and strong hospital effects. The third, largest cluster includes older but less comorbid patients, where pneumonia and hospital variability both significantly impact survival.

Discussion

Results show that respiratory complications and hospital effects vary across patient profiles. Hospital-level heterogeneity is most pronounced among clinically vulnerable but less comorbid patients, suggesting that healthcare context influences survival in specific subgroups. The proposed multilevel survival CWM offers a flexible parametric framework for clustering hierarchical time-to-event data while disentangling patient- and group-level effects.

Linked shrinkage for interaction modelling in time-to-event data

15:45-16:00

Presenter: Jeroen Hoogland

Authors: Jeroen Hoogland; Mark van de Wiel

Affiliations: Department of Epidemiology and Data Science, Amsterdam Medical Centers, the Netherlands, Amsterdam

Introduction

Modeling two-way interactions in time-to-event data may offer a flexible yet interpretable alternative to black-box machine learning approaches in settings of moderate dimension. Previous work suggests the use of a linked-shrinkage prior to cope with the quadratically increasing dimension of the interaction space, linking interaction parameters to their corresponding main effects. We aim to develop a parametric survival model with a linked-shrinkage prior that enables efficient estimation and interpretable inference in moderately high-dimensional settings.

Methods

Extending our earlier framework for interaction modeling, we propose a Bayesian local shrinkage approach in which the shrinkage of interaction effects is adaptively linked to that of their corresponding main effects. This linkage promotes parsimonious models while retaining flexibility, allowing key interactions to emerge naturally when supported by the data. The method accommodates typical time-to-event data challenges such as left truncation, right censoring, and interval censoring within

a parametric framework, and is implemented in RStan. In terms of interpretation, efficient computations of feature and interaction Shapley values are shown to quantify individual-specific variable importance and associated uncertainty.

Results

Simulation studies indicate that the proposed linked-shrinkage survival model achieves accurate parameter estimation and competitive predictive performance in moderate-dimensional scenarios. The results suggest that the proposed method can approach the predictive ability of more complex machine learning methods while preserving interpretability and parameter inference.

Discussion

The proposed model provides an attractive middle ground between traditional regression and machine learning approaches for survival analysis in problems of moderate dimension that are typical in medical applications. Combining interpretability and inference while retaining flexibility make it a broadly applicable tool for a wide range of time-to-event analyses.

Calculating Updated Prognosis for Longer-Term Survivors: Approaches to Estimating Conditional Survival in the Presence of Time-Dependent Covariate Effects

16:00-16:15

Presenter: Aiden Smith

Authors: Aiden Smith [1]; Elisavet Syriopoulou [2]; Therese Andersson [2]; Paul Lambert [3, 2]; Mark Rutherford [1]

Affiliations: [1] University of Leicester, Biostatistics Research Group, Division of Public Health and Epidemiology, School of Medical Sciences, Leicester; [2] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm; [3] Cancer Registry of Norway, Norwegian Institute of Public Health, Oslo

Introduction

Prognostic survival estimates have historically anchored on 'time zero' (diagnosis), failing to reflect a patient's evolving risk profile over time. Conditional survival provides a dynamic update of prognosis, yet its estimation often relies on standard survival approaches, such as the Cox model, which assume Proportional Hazards (PH). In real-world settings, biological factors often exhibit time-varying effects, violating the PH assumption. Failing to account for Non-Proportional Hazards (NPH) risks producing biased estimates that are subsequently used in clinical decision-making, which may be detrimental to patient care. The predictive accuracy of model-based methods which account for NPH, such as Flexible Parametric Models (FPMs), are subject to appropriate specification of time-dependent effects in the model itself. We propose a Pseudo-observation (PSO) framework as a pragmatic alternative to estimate conditional survival in prognostic settings.

Methods

We compare various approaches for handing time-varying covariate effects and evaluate the impact of relaxing the proportional hazards assumption on the estimation of conditional survival. We include a 'Stacked Pseudo-Observations' (PSO-Stack) approach, where the pseudo-observations for each individual at multiple timepoints are modelled using a single unified Generalized Linear Model (GLM) with time-covariate interactions. We conducted a simulation study comparing PSO approaches against model-based estimation using FPMs, across varying sample sizes. Performance was evaluated via bias and coverage probabilities relative to a large-sample non-parametric 'truth' under both PH and NPH data-generating mechanisms. We exemplify our approach using SEER colorectal cancer data.

Results

In scenarios involving NPH, standard PH models produced spurious results by assuming a constant hazard ratio over time. Conversely, the PSO-Stack approach accurately approximated the non-parametric truth without requiring specification of the underlying hazard function. Furthermore, the PSO method avoided convergence issues associated with delayed-entry models and complex FPMs in smaller sample sizes.

Discussion

The PSO-Stack approach offers a computationally efficient, unified method for estimating unconditional and conditional survival while accounting for time-dependent predictors. By removing reliance on the restrictive PH assumption, this method mitigates the risk of statistical over-simplification. This ensures accurate estimation of conditional survival in scenarios where NPH are observed and removes the need to directly define time-dependent effects in the modelling process.

Benchmarking multi-step methods for the dynamic prediction of survival with numerous longitudinal predictors

16:15-16:30

Presenter/Author: Mirko Signorelli

Affiliations: Leiden University, Leiden

Introduction

Collecting longitudinal health information has become more and more common thanks to the advancement of technologies to measure biometrical outcomes and to store and process data. These technological advancements have led to an increasing availability of repeated measurement data that can be used to monitor a subject's health status and disease progression, making it possible to update predictions of the probability to experience an adverse event each time a new measurement is taken. However, until recently there was a lack of statistical methods that could handle many longitudinal covariates as predictors of survival outcomes.

Methods

Recently, the increasing availability of repeated measurements in biomedical studies has fostered the development of several new statistical methods for the dynamic prediction of survival outcomes in settings where a large, potentially high-dimensional number of longitudinal covariates is available. These methods differ in the approaches they use to model and summarize the trajectories of the longitudinal covariates (multivariate functional PCA or mixed-effects models), and in the way they specify the relationship between the longitudinal covariates and the survival outcome of interest (Cox model or random survival forests). Because these methods are still quite new, little is known about their applicability, limitations and performance when applied to real-world data.

Results

We review the aforementioned methodological developments, and discuss how the different modelling choices that they make can have an impact on the possibility to accommodate unbalanced study designs, and on computing time. Furthermore, we present the results of a systematic comparison of the predictive performance of these methods based on several real-world datasets that differ in terms of type of outcome, sample size, number of longitudinal covariates, and length of the follow-up.

Discussion

We present an overview of the advantages and limitations of each method, including methodological considerations as well as practical aspects such as software availability and computing time. Finally, we discuss possible avenues for future methodological developments.

Reference: Signorelli, M., Retif, S. (2025). Benchmarking multi-step methods for the dynamic prediction of survival with numerous longitudinal predictors. The International Journal of Biostatistics. DOI: https://doi.org/10.1515/ijb-2025-0049

Locally Estimated Scatterplot Smoothing for Censored Survival Data

16:30-16:45

Presenter: David van Klaveren

Authors: David van Klaveren [1, 2]; Peter C. Austin [3, 4, 5]; Patrick W. Serruys [6]; Ben Van Calster [7, 8, 9]; Frank E. Harrel Jr. [10]

Affiliations: [1] Department of Public Health, Erasmus MC University Medical Center, Rotterdam; [2] Predictive Analytics and Comparative Effectiveness Center, Tufts Medical Center, Boston; [3] ICES, Toronto; [4] Institute for Health Policy, Management and Evaluation, University of Toronto, Toronto; [5] Sunnybrook Research Institute, Toronto; [6] CORRIB Research Centre for Advanced Imaging and Core Laboratory, University of Galway, Galway; [7] Department of Development and Regeneration, KU Leuven, Leuven; [8] Leuven Unit for Health Technology Assessment Research (LUHTAR), KU Leuven, Leuven; [9] Julius Center for Health Sciences and Primary Care, University

Medical Center Utrecht, Utrecht; [10] Department of Biostatistics, Vanderbilt University School of Medicine, Nashville

Introduction

In the absence of censoring, Locally Estimated Scatterplot Smoothing (LOESS) is a powerful tool to estimate the relationship between a continuous variable and the probability of an event and is often used to visually assess the calibration of prognostic models. We aimed to develop a nonparametric estimator of the relationship between a continuous variable and the probability of an event within a specified time horizon in the presence of right-censoring.

Methods

We exploited the insight that the nonparametric Kaplan-Meier estimator can be written as a weighted average of the event indicators, with weights equal to the inverse probability of observing the event. These weights can easily be estimated by the inverse Kaplan-Meier estimator or - when censoring is likely to depend on the continuous variable - by a Cox regression with a spline function of the variable. The central idea of our method is that we can directly apply standard LOESS to the weighted event indicators. We used the resulting LOESS for censored data (LOESScd) to estimate the association between continuous predictors and the probability of death within 10 years for patients with complex coronary artery disease in the SYNTAXES trial. We examined the statistical properties of LOESScd in simulations (2,000 samples of 1,800 patients) motivated by four typical validation scenarios of the prognostic SYNTAX SCORE 2020.

Results

LOESScd estimates closely followed grouped Kaplan-Meier estimates, especially in case of weak smoothing. With stronger smoothing the LOESScd provided more stable estimates. In simulated data of all validation scenarios (perfect calibration overestimation linear overfitting non-linear mis-calibration), the LOESScd-based calibration curves were on average close to the true calibration curves of 10-year mortality predictions according to the SYNTAX SCORE 2020. Bias was limited and coverage was close to nominal, both when censoring was random and when the probability of censoring increased over the full range of predicted 10-year mortality.

Discussion

LOESScd is a promising nonparametric estimator of the relationship of a continuous variable with the probability of an event in the presence of right-censoring. LOESScd is especially useful for estimating calibration curves when validating prognostic models and avoids arbitrary binning of predicted values.

A Strictly Proper Metric to Evaluate Dynamic Predictions for Left-Truncated and Right-Censored Data with Asymmetric Misclassification

16:45-17:00

This is one of this year's ISCB StCA award-winning abstracts.

Presenter: Arnau Garcia-Fernández

Authors: Arnau Garcia-Fernández [1, 2]; Dimitris Rizopoulos [1]; Judith Kooiman [2]; Pedro Miranda-Afonso [1]

Affiliations: [1] Department of Epidemiology and Biostatistics, Erasmus University Medical Center, Rotterdam; [2] Department of Obstetrics and Gynecology, Erasmus University Medical Center, Rotterdam

Introduction

In high-risk pregnancies, antenatal corticosteroids are most effective within one week before birth, so accurately predicting imminent delivery is important. Failing to treat a patient who will deliver soon (false negative) is more serious than treating too early (false positive), and delayed cohort entry causes left truncation. These features challenge available metrics for evaluating predictions.

Strictly proper metrics, such as the Brier score (BS), ignore the asymmetry in the consequences of false positives and negatives. Clinical utility metrics, such as net benefit (NB) and expected cost, are not strictly proper and rely on a single decision threshold, for which consensus can be difficult to achieve. Decision curves display these metrics across multiple thresholds but do not provide a single summary value to compare models. Finally, existing estimators can be biased under left truncation.

Methods

To combine strict properness and clinical utility without relying on a single threshold, we propose the integrated weighted expected cost (IWEC). This metric integrates the expected cost over all thresholds using a clinically informed weight function. We show that every centered strictly proper scoring rule admits an IWEC representation. We then propose an inverse-probability-weighting estimator for left-truncated and right-censored (LTRC) data, proving its consistency and asymptotic normality. Finally, we introduce a weighted decision curve whose integral corresponds to the IWEC.

Results

We conducted a simulation study to compare the new metric with established alternatives. BS and area under the receiver operating characteristic curve favored models with strong statistical discrimination, whereas NB and IWEC favored those with greater clinical value. Although NB was sensitive to the choice of threshold, IWEC was robust to threshold misspecification.

We then applied the metrics to the motivating dataset, using weighted decision curves to assess the clinical usefulness of alternative shared-parameter joint models, with the clinical relevance of each threshold weighted by domain experts.

Discussion

Our results suggest that model evaluation may differ substantially depending on whether emphasis is placed on statistical discrimination or clinical usefulness. IWEC provides a summary measure that reflects clinical priorities without requiring selection of a single decision threshold. The proposed methods are available in the JMbayes2 R package.

Tue 29.09.2026

11:00–12:30

LLMs for Clinical Text and Data

Tue, 29.09.26, 11:00–12:30, Room: K2

Chair: Phillip Richter-Pechanski

LLM-Based Data Augmentation for Clinical Documents: A Class-Level Analysis

11:00-11:15

Presenter: Vittorio Torri

Authors: Vittorio Torri [1]; Francesca Ieva [1, 2]

Affiliations: [1] MOX - Modelling and Scientific Computing Lab, Department of Mathematics, Politecnico di Milano, Milano; [2] HDS - Health Data Science Centre, Human Technopole, Milano

Introduction

Clinical documents are a key source of information within electronic health records. However, supervised models for clinical text classification are limited by the scarcity of annotated data and class imbalance, especially for rare conditions. Synthetic data augmentation using Large Language Models (LLMs) has been proposed to address these limitations, but it is unclear whether it consistently improves classification across clinical categories or introduces biases affecting downstream use. We evaluate LLM-based augmentation in clinical document classification, focusing on class-level behaviour and correlations with real and synthetic data variability.

Methods

We consider two classification tasks on Italian clinical documents with multiple classes and strong imbalance: E3C case reports (2,040 documents, 14 classes), and Emergency Department notes from San Giovanni Bosco Hospital (Turin) (2,957 documents, 8 classes), annotated for disease categories. Training data are augmented with synthetic documents generated conditionally on class labels using two Italian LLMs (LLaMAntino-2 and LLaMAntino-3), with different prompting and sampling strategies. Synthetic samples are progressively added after filtering low-quality outputs. Performance is evaluated via stratified cross-validation using precision, recall, and F1-score. We quantify variance in real and synthetic documents using lexical and embedding-based measures and assess its association with augmentation effects and sampling hyperparameters via Spearman correlation.

Results

In both datasets, augmentation shows heterogeneous effects across clinical categories. While some underrepresented conditions benefit (ΔF1 up to +5%), others experience performance degradation. Precision gains (up to +10%) are often larger than recall improvements (up to +5%), with a frequent trade-off. LLaMAntino-3 provides more consistent improvements than LLaMAntino-2. We observe a negative correlation between performance gains and document variability, particularly for synthetic data (ρ up to -0.59$\pm$0.30). The relationship between sampling temperature and variance is moderate (ρ up to 0.61$\pm$0.14).

Discussion

Synthetic data augmentation can improve clinical document classification but requires careful control to avoid deviations from real data distributions. Our results indicate that performance is closely linked to the variance of generated data, with moderate dispersion associated with better gains. This suggests that effective augmentation should prioritize alignment with real clinical language rather than diversity alone. The consistency across two datasets supports the robustness of these findings, although further validation is needed.

Evaluating Large Language Models for Structuring Free-Text Patient Experience Comments into Operationally Relevant Signals

11:15-11:30

Presenter: Esra Tokur Sonuvar

Authors: Esra Tokur Sonuvar; Sabrina Friebe; Toralf Kirsten

Affiliations: Institut für Medizinische Informatik, Statistik und Epidemiologie, Universität Leipzig, Leipzig

Introduction

Free-text patient feedback often captures problems that are not fully reflected in structured survey items, but using these comments in routine analysis remains difficult. This study examined whether large language models (LLMs) can organize NHS general practice PREM comments into a structured annotation scheme while preserving the information needed for service evaluation..

Methods

A random sample of 500 comments was drawn from a publicly accessible dataset of 1,952 NHS GP patient feedback records from England. Two expert annotators independently labeled each comment for primary issue category, severity, urgency, evidence span, and suggested follow-up action. Disagreements were resolved through discussion to create a final reference standard. Claude (claude-sonnet-4-6) and an OpenAI model configured as gpt-5.2 were then used in a zero-shot setting to generate structured JSON outputs with the same schema. Agreement with the reference

standard was assessed using accuracy and Cohen's kappa with 95% bootstrap confidence intervals. Rule-based validation covered JSON validity, schema compliance, allowed-value compliance, evidence grounding, and evidence length.

Results

Inter-expert agreement was substantial across all fields, with kappa values of 0.811 for primary category, 0.815 for severity, and 0.804 for urgency. Claude performed better than GPT across all evaluated fields. For the primary issue category, Claude achieved an accuracy of 0.892 (95% CI: 0.864-0.918) and a kappa of 0.863 (95% CI: 0.828-0.897), whereas GPT reached 0.736 and 0.669, respectively. The difference was especially marked for urgency, where Claude achieved a kappa of 0.756 compared with 0.480 for GPT. Both models achieved 100% JSON validity, 100% schema compliance, and 100% allowed-value compliance. Evidence grounding was higher for Claude than for GPT (89.8% vs. 80.0%), and GPT produced 4.0% overlength evidence spans whereas Claude produced none.

Discussion

These results show that LLMs can be used to turn free-text PREM comments into a consistent structured format, although model quality differed clearly. Claude was closer to expert annotation and handled evidence extraction more reliably. At the same time, categories with overlapping meanings remained more difficult, which suggests that human oversight remains important, particularly for comments with direct service or organizational relevance.

Embedding-Based Evaluation and Benchmarking Framework for Optimizing LLM Post-Processing of Medical Transcriptions from a Local Whisper System

11:30-11:45

Presenter: Ahmad Wattar

Authors: Ahmad Wattar [1]; Jan Christoph [2]; Christoph Demus [1]; Patrick Liegel [1]; Christian Jäger [1]

Affiliations: [1] Research Group (Bio-) Medical Data Science, Faculty of Medicine, Martin Luther UniversityHalle-Wittenberg, Halle (Saale), Germany, Halle (Saale); [2] Research Group (Bio-) Medical Data Science, Faculty of Medicine, Martin Luther UniversityHalle-Wittenberg, Halle (Saale), Germany, Halle

Introduction

Locally deployed speech-to-text systems such as Whisper enable privacy-preserving transcription of medical encounters. However, the resulting transcripts are often lengthy and unstructured, making them difficult to integrate directly into clinical documentation workflows. Large language models (LLMs) offer a promising approach for transforming transcripts into concise clinical summaries. Evaluating the quality and reliability of these summaries remains challenging, particularly in the absence of gold-standard reference summaries.

Methods

We propose a reference-free evaluation framework for benchmarking LLM-based post-processing of clinical transcripts generated by a locally deployed Whisper system. The framework assesses summary quality using embedding-based semantic similarity between the source transcript and the generated summary. In addition, it evaluates generation stability across repeated runs and complements automatic metrics with structured human evaluation. The framework was applied to 26 German physician-patient conversations. Four LLMs generated summaries using standardized prompts and decoding parameters, with four runs per transcript, resulting in 416 summaries in total. A subset of summaries was additionally evaluated by three human raters with respect to factual consistency, completeness, and clinical usefulness.

Results

Across embedding models, GPT-OSS-120B and MedGemma-27B consistently achieved higher transcript-summary similarity scores than the other evaluated LLMs. While absolute similarity values varied across embedding models, the relative ranking of summarization models remained largely consistent, indicating robustness of the evaluation framework. Stability analysis further showed high consistency of generated summaries across repeated runs, with cosine similarities typically exceeding 0.90. Temperature experiments demonstrated decreasing semantic similarity at higher sampling temperatures. In addition, preliminary human evaluation revealed partial agreement between embedding-based similarity signals and human judgments of summary quality.

Discussion

The proposed framework enables scalable and reference-free evaluation of LLM-based clinical summarization in privacy-sensitive environments. By combining embedding-based semantic similarity, stability analysis, and human evaluation, it supports systematic benchmarking and optimization of summarization models without requiring reference summaries. The results indicate that embedding-based metrics provide reliable signals for relative model comparison and can support the integration of LLM-based post-processing into local speech-to-text pipelines such as Whisper.

The full paper of this abstract will be published in the GMDS series "Studies in Health Technology and Informatics" with doi: 10.3233/SHTI261001, URL:https://dx.doi.org/10.3233/SHTI261001

Large Language Model-Based Text-to-Cypher Generation for Question-Answering on Biomedical Knowledge Graphs

11:45-12:00

Presenter: Adem Abdelmoula

Authors: Suteera Seeha [1]; Adem Abdelmoula [2]; Martin Boeker [3]; Luise Modersohn [4]; Peter Pallaoro [5]

Affiliations: [1] Chair of Medical Informatics, Institute of AI and Informatics in Medicine (AIIM), TUM University Hospital, Technical University of Munich, Munich, Germany, Munich; [2] Technische Universität München (TUM), Munich; [3] Technische Universität München (TUM), München; [4] Institute for Artificial Intelligence and Informatics in Medicine, Chair of Medical Informatics, Medical Center rechts der Isar, Technical University of Munich, Munich, Germany, München; [5] Institute for AI and Informatics in Medicine (AIIM), Klinikum rechts der Isar, School of Medicine, Technical University of Munich, Munich

Introduction

Knowledge graphs are widely adopted due to their flexible structure and ability to represent large-scale, heterogeneous data. However, as these graphs grow, the queries used to retrieve information become more complex, making efficient information retrieval and question answering more challenging.

Methods

This work investigates (1) the impact of few-shot versus zero-shot prompting on Cypher query generation, and (2) performance differences between small (~7B) and mid-size (~32B) large language models (LLMs) on both of these approaches. We propose a system that translates natural language questions into Cypher queries for execution on a Neo4j knowledge graph and evaluates answer quality on the PrimeKG and BioHopR datasets. To address data protection and cost constraints, we focus on locally deployable LLMs that require no additional training, using zero-shot and few-shot prompting strategies.

Results

32B models generate syntactically valid Cypher queries using only the knowledge graph schema information. Few-shot prompting with 4 examples improves F1 Score by 61% compared to zero-shot prompting. While 7B models struggle to generate syntactically valid queries in the zero-shot setting, few-shot prompting substantially improves their performance, enabling high query syntax validity and F1 scores above 0.8.

Discussion

Few-shot prompting heavily improves query generation quality and enables effective use of smaller models, supporting the practical deployment of locally hosted LLMs for knowledge graph question answering.

The full paper of this abstract will be published in the GMDS series "Studies in Health Technology and Informatics" with doi: 10.3233/SHTI261000, URL:https://dx.doi.org/10.3233/SHTI261000

Probing Clinical Reasoning in Large Language Models via Triplet Comparisons

12:00-12:15

Presenter: Fabian Kabus

Authors: Fabian Kabus [1]; Amal Abed [1]; Daiana Stolz [2]; Heinz Wiendl [3]; Hannah Bast [4]; Harald Binder [5]

Affiliations: [1] Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Germany, Freiburg im Breisgau; [2] Clinic of Pneumology, Medical Center - University of Freiburg, Freiburg im Breisgau; [3] Klinik für Neurologie und Neurophysiologie - Universitätsklinikum Freiburg, Freiburg; [4] Department of Computer Science, University of Freiburg, Germany, Freiburg im Breisgau; [5] Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Germany, Freiburg

Introduction

Large language models are increasingly deployed in clinical decision-support systems, yet their reasoning processes remain opaque.

Understanding which combination of clinical variables drive model decisions is essential for safe deployment.

We present a statistical framework to probe LLM behavior using structured triplet comparison experiments, where models judge which of two patients is more similar to a reference case based on clinical prognosis.

Methods

We construct triplet questions from patient cohorts characterized by multiple clinical variables.

For each triplet, we compute a delta encoding that captures how the reference patient is positioned relative to the comparison patients across all variables.

Responses are discretized into contingency tables and analyzed using hierarchical log-linear models to identify variable interactions influencing model decisions.

Since exhaustive triplet evaluation is combinatorially infeasible, we propose an adaptive acquisition strategy.

The method employs two-stage candidate selection: first screening triplets by predictive entropy and cell-occupancy novelty, then evaluating shortlisted candidates via structure-aware lookahead that simulates how each query would affect the inferred interaction structure.

Triplets are selected to maximize expected structural change, enabling efficient discovery under limited query budgets.

Results

We applied this framework to compare LLM configurations across model size, prompt design, and retrieval augmentation.

Log-linear model selection revealed distinct variable interaction structures across configurations.

Prompt variations - including syntactic variation, description detail, and response format - produced detectable differences in decision behavior.

Cross-model comparisons between larger and smaller instruction-tuned models showed significant divergence in how clinical variables influenced similarity judgments.

Discussion

This framework provides a principled approach to characterizing LLM clinical reasoning without requiring internal model access.

The adaptive acquisition strategy substantially reduces query requirements while preserving structural discovery.

By revealing which variable combinations drive model decisions, this method enables systematic comparison of LLM configurations and can inform model selection and prompt engineering for clinical applications.

Privacy Assessment of Large Language Models Following Instruction-Based Fine-Tuning on Medical Reports

12:15-12:30

Presenter: Stefan Lenz

Authors: Stefan Lenz [1]; Maximilian Wehle [1]; Miguel Mioskowski [1]; Verena Kauth [2]; Torsten Panholzer [1]; Gregor Duwe [2]

Affiliations: [1] Institute of Medical Biostatistics, Epidemiology and Informatics, University Medical Center of the Johannes Gutenberg University Mainz, Mainz; [2] Department of Urology and Pediatric Urology, University Medical Center of the Johannes Gutenberg University Mainz, Mainz

Introduction

Sharing Large Language Models (LLMs) that have been fine-tuned on clinical data raises concerns regarding the potential memorization of sensitive patient information. Yet, systematic evaluation of privacy-related vulnerabilities in clinically fine-tuned LLMs has remained limited. This study evaluates two attack scenarios: inferring clinical information from patient names and re-identifying individuals from redacted documents.

Methods

We instruction-tuned MedGemma-1.5-4b-it with QLoRA to analyze patient documentation for urological tumor boards and answer specific questions regarding applied therapies. Training was conducted for one epoch on a balanced dataset of 8,664 questions with associated clinical context (50%/50% yes/no answers). To optimize the logical derivation of complex positive findings, we employed reasoning distillation using a teacher-student approach (Qwen-3 8B/32B). We injected 12 personal names from German and Turkish cultural backgrounds into ~1.0% of the training data to simulate a scenario of attempted but incomplete anonymization. We then inquired details of the medical history using the injected names via prompts and attempted to reconstruct names via generative text continuation.

Results

Instruction-tuning significantly enhanced overall model performance, leading to substantial improvements in classification accuracy, strict format adherence, and highly accurate citations for positive predictions. When querying details about the medical history for injected names, the fine-tuned model achieved only 17% accuracy compared to a baseline of 8%. While the relative risk doubled, true positives remained constant. Evaluating the output probabilities revealed a general bias towards predicting 'Yes' rather than context-specific memorization. Text continuation for complete name reconstruction yielded zero completions. When provided with first names, both the fine-tuned and base models reconstructed the same 2 of 12 names with similarly low probabilities.

Discussion

The findings primarily reflect model and training biases rather than leakage of personal data during fine-tuning. This suggests that fine-tuning on clinical data does not necessarily result in privacy risks even if the training data has not been perfectly anonymized. Future work should investigate and combine additional methods such as those applied in this study to develop a systematic framework for secure model sharing.

Clinical Trajectories and Events

Tue, 29.09.26, 11:00–12:30, Room: K3

Chair: Francesca Ieva

Typical Healthcare Pathways as a Basis for Admixture Modeling of Patient Trajectories

11:00-11:15

Presenter: Maryam Farhadizadeh

Authors: Maryam Farhadizadeh [1, 2]; Carola S. Heinzel [3]; August Sigle [4]; Harald Binder [5, 2]; Frederik Wenz [6]; Jan Hasenauer [7, 8]; Peter Pfaffelhuber [3]; Nadine Binder [1, 2]

Affiliations: [1] Institute of General Practice/Family Medicine, Medical Faculty and Medical Center, University of Freiburg, Freiburg im Breisgau; [2] Freiburg Center for Data Analysis, Modeling and AI, University of Freiburg, Freiburg im Breisgau; [3] Department of Mathematical Stochastics, University of Freiburg, Freiburg im Breisgau; [4] Department of Urology, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg im Breisgau; [5] Institute of Medical Biometry and Statistics (IMBI), Medical Faculty and Medical Center, University of Freiburg, Freiburg; [6] Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Freiburg im Breisgau; [7] University of Bonn, Bonn; [8] Life & Medical Sciences Institute, University of Bonn, Bonn

Introduction

Healthcare pathways derived from routine care data are inherently heterogeneous and temporally complex, reflecting diverse patient trajectories and variations in care delivery. This complexity poses significant challenges for identifying meaningful patient subgroups using standard clustering approaches, which often fail to capture the underlying structure of care processes.

Methods

To address these challenges, we propose a framework that explicitly separates cohort-level pathway identification from patient-level inference. At the cohort level, we develop a rule-based algorithm to derive a simplified and interpretable representation of care processes by identifying typical healthcare pathways, summarized as a compact pathway graph. These pathways are subsequently modeled as Markov chains, capturing transitions between clinical events. The resulting pathway-specific models are then incorporated as structured components within an admixture model. This approach allows each patient trajectory to be represented as a probabilistic mixture of typical pathways, rather than being assigned to a single pathway component, thereby accommodating heterogeneity in patient-level processes.

Results

The inferred admixture weights provide a compact and low-dimensional representation of patient trajectories, enabling downstream analyses such as subgroup identification and characterization. Stability analyses conducted across multiple train-test splits demonstrate consistent pathway structures and robust patient-level mixture representations, supporting the reliability of the proposed framework. When applied to routine care data from prostate cancer patients undergoing radical prostatectomy, the method identifies clinically plausible and interpretable care patterns. Furthermore, it reveals distinct subgroups of patients with similar sequences of coded clinical events, highlighting its ability to capture meaningful variations in treatment pathways.

Discussion

This framework provides a stable, interpretable, and flexible approach for simplifying complex healthcare pathways while preserving essential structural information. By enabling a probabilistic representation of patient trajectories, it facilitates the identification and characterization of clinically meaningful patient subgroups in real-world healthcare data.

Clustering Multivariate Categorical Sequences Under Censoring and Truncation: A Benchmark and Transformer-Based Approach for Clinical Trajectory Analysis

11:15-11:30

Presenter: Ottavio Khalifa

Authors: Ottavio Khalifa [1]; Viet-Thi Tran [2, 3]; François Petit [4]

Affiliations: [1] Université Paris Cité, Université Sorbonne Paris Nord, INSERM, INRAE, Centre for Research in Epidemiology and StatisticS (CRESS), Paris, France., Paris; [2] Université Paris-Cité, Center of Research in Epidemiology and StatisticS, METHODS Team, Paris; [3] Centre D'Epidemiologie Clinique, AP-HP, Hôpital Hôtel Dieu, F-75004, Paris, France, Paris; [4] Université Paris Cité, Université Sorbonne Paris Nord, INSERM, INRAE, Centre for Research in Epidemiology and StatisticS (CRESS), Paris, France, Paris

Introduction

Identifying patient subgroups from longitudinal trajectories is a core task in public health and epidemiology. In practice, these sequences are multivariate, right-censored, left-truncated, have missing data, and accompanied by time-fixed covariates. A scoping review of 129 clustering methods for categorical time series [1] identified a gap: no existing method jointly handles these characteristics at scale. We present a benchmark and a transformer-based framework targeting this gap, with application to the Compare long COVID cohort [2].

Methods

The benchmark generates multivariate sequences from mixture models of stationary HMMs and non-stationary GLMs. Baselines are EM-fitted mixture HMMs and Optimal Matching with TRATE costs [3], combined with agglomerative and k-medoids clustering. To mimic left truncation and right censoring, we evaluate on interior windows - random contiguous slices of each sequence - rather than full sequences.

The approach proceeds in two stages. A bidirectional autoregressive transformer with relative positional encoding learns representations via self-supervised training, with prediction of next token as the core task. Each sequence is augmented into two independently censored views, forming a contrastive pair. This censoring augmentation accepts any survival mechanism, including one estimated from baseline covariates, allowing the model to account for informative censoring. Clustering is then applied to the learned latent space. Unlike Optimal Matching, the approach scales linearly with dataset size.

Results

The transformer outperforms both baselines in Adjusted Rand Index (ARI), with largest gains under non-stationarity and short sequences. Contrastive augmentation is the main driver under truncation tail-weighted pooling proves more robust than standard alternatives.

Discussion

This framework addresses a gap in the existing literature on categorical time series clustering. Application to the Compare long COVID cohort will test clinical interpretability of clusters under real-world follow-up conditions.

[1] Khalifa O et al. Clustering methods for Categorical Time Series and Sequences: A scoping review. arXiv:2509.07885. 2025.

[2] Servier C et al. Trajectories of the evolution of post-COVID-19 condition, up to two years after symptoms onset. Int J Infect Dis. 2023133:67-74.

[3] Gabadinho A et al. Analyzing and Visualizing State Sequences in R with TraMineR. Journal of Statistical Software. 201140(4):1-37.

Bayesian Profile Regression using Variational Inference for Population-Scale Data: Application to Clustering Multiple Long-Term Conditions Conditioning on Mortality

11:30-11:45

Presenter: Rhiannon K Owen

Authors: James Rafferty [1]; Keith Abrams [2]; Munir Pirmohamed [3]; Mark Davies [4]; Rhiannon K Owen [1]

Affiliations: [1] Swansea University, Swansea; [2] University of Warwick, Coventry; [3] University of Liverpool, Liverpool; [4] Cardiff University, Cardiff

Introduction

Multiple long-term conditions (MLTC) are increasingly observed in clinical practice globally and cluster in many different patterns. Existing clustering methods for population scale datasets (e.g. latent class analysis, k-means clustering) do not condition on an outcome of interest such as mortality.

Methods

Bayesian Profile Regression (BPR) is a model that combines a Dirichlet Process Mixture model (the mixture component) with a hierarchical regression model (the response component) to form clusters conditional on covariates and an outcome. We developed a BPR model and compared fits using full-rank Stochastic Variational Inference (SVI) with the No-U-Turn (NUTS) sampler in simulation studies. We used BPR with SVI to identify clusters of MLTC in population-scale data held in the Secure Anonymised Information Linkage (SAIL) databank, using the Elixhauser comorbidity index for disease definitions. The model was conditioned on mortality as the outcome, and adjusted for age at cohort entry, socioeconomic deprivation, and sex.

Results

Biases in both mixture and response components were small and centred on zero for both fitting methods. Coverage for the for the mixture component using SVI compared well with NUTS. The response component coverage using SVI was smaller than NUTS fits. Our SAIL cohort contained 1.3 million individuals and found 33 clusters. Notable clusters included seven that featured high probabilities of cancer and metastatic cancer (cluster probability range: [<0.001, 0.011], mortality probability range: [0.217, 1.000], Metastatic cancer probability range: [0.667, 0.987]).

Discussion

Clusters featuring metastatic cancer and cardiovascular diseases, such as congestive heart failure, were most strongly associated with a high probability of mortality. The separate clusters featuring cancer were likely to represent different groups of patients, for example i) treatable early-stage disease, ii) late-stage disease with limited treatment options, iii) comorbidities that modified their mortality risk and iv) comorbidities induced by cancer treatment. Without conditioning on an outcome of interest, existing methods would not identify these clusters within diseased groups. We have shown that SVI can be a useful and accurate method for fitting Bayesian Profile Regression to large-scale datasets such as routinely collected electronic health records where Monte Carlo methods are prohibitively time consuming or intractable.

NexT-ICU: A resource efficient time-series foundation model for critical care using selective state space representations

11:45-12:00

Presenter: Aditya Kumar

Authors: Aditya Kumar [1, 2]; Ivan Lukic [2]; Mario A. Cypko [1, 2]; Oliver Amft [1, 2]

Affiliations: [1] Hahn-Schickard, Freiburg; [2] Intelligent Embedded Systems Lab, University of Freiburg, Freiburg

Introduction

Critical care is characterised by heterogeneous patients, irregularly sampled physiological data, and multiple clinical objectives. Traditional task-specific models struggle to generalise across data variability originating from patients and clinical settings. In contrast, foundation models generalise to diverse downstream tasks, but increase resource requirements. State-of-the-art approaches, e.g., STraTS [1], rely on attention mechanisms with quadratic time and memory complexity. We propose NexT-ICU, a foundation model based on selective state space representations that (1) achieves linear-time scalability and (2) supports multiple prediction tasks.

Methods

We utilised a 5-layer unidirectional Mamba-based encoder to model temporal dynamics. Observations were encoded using embeddings for variable identity, measurement value, and temporal context (Δt), the latter to explicitly handle irregularly sampled continuous data. Static patient information was integrated via feature-level fusion. To learn task-agnostic representations, the model training was self-supervised (contrastive learning and masked value modelling) on >200,000 eICU and >95,000 MIMIC-IV ICU stays. We evaluated NexT-ICU against STraTS on downstream classification (mortality, sepsis) and regression (kidney function, length of stay).

Results

In preliminary results on MIMIC-IV, NexT-ICU performed similarly to STraTS across all tasks. NexT-ICU achieved higher AUROC for both mortality (0.878 ± 0.000 vs 0.830 ± 0.014) and sepsis prediction (0.828 ± 0.020 vs 0.748 ± 0.004). NexT-ICU reduced MAE for kidney function (0.255 ± 0.006 vs 0.264 ± 0.013 mg/dL) and for length of stay (36.2 ± 0.3 vs 37.8 ± 0.3 hours). For NexT-ICU, we reached peak GPU memory of ~1.5 GB (vs >40 GB at sequence length 8000) and higher throughput (~250 vs ~10 samples), while model sizes were similar (~0.26M parameters).

Discussion

NexT-ICU achieves linear time and memory complexity by utilising a selective state space representations, which highlights its potential for practical and resource-efficient deployment in clinical applications. Furthermore, NexT-ICU can handle long and irregular patient trajectories. As a foundation model trained on large heterogeneous ICU cohorts, NexT-ICU learns generalisable representations. Ongoing work shall expand the evaluation framework to strengthen the robustness and clinical relevance of the model.

[1] Tipirneni, S. and Reddy, C.K., 2022. Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. ACM Transactions on Knowledge Discovery from Data, 16(6), pp.1-17.

Augmented Time-to-Event methods for Events Defined by Continuous Measurements

12:00-12:15

Presenter: Yingjuan Zhang

Authors: Yingjuan Zhang; James Wason

Affiliations: Population Health Sciences Institute, Newcastle University, UK, Newcastle Upon Tyne

Introduction

In some time-to-event settings, the event is defined by some continuous measurement(s) taking a range of values. For example, in rheumatoid arthritis, flare is defined as continuous measure called DAS28-CRP taking values above 3.2. Lin and Wason [1] proposed a time-to-event augmented binary method for multiple endpoints, which incorporates the continuous variable directly rather than dichotomizing it, improving analysis efficiency. While their approach primarily addresses combined endpoints consisting of binary and continuous variables, we improve the method for settings with continuous endpoints only.

Methods

In this work, we consider a discrete-time first-hitting-time formulation to incorporate event information when only the continuous variable is available. For censored subjects, the joint probability that biomarker values remain below threshold c across all measurements is computed through multivariate integration. For subjects experiencing an event, the probability is obtained as the difference between the cumulative probability up to pre-event and event times, thereby capturing the transition to event occurrence within the discrete measurement framework. The method allows inference on effect measures such as the hazard ratio, but better incorporates continuous information to improve statistical efficiency.

Results

Simulation studies demonstrate that the extended method is consistent and robust, with improved power compared to standard methods across diverse settings. As a case study, we considered the Bio-Flare study, which investigates the biological processes that trigger rheumatoid arthritis flares using DAS28-CRP. The threshold is set to c = 3.2. We explore the estimated average hazard ratio measuring association between sex and methotrexate use as separate predictors for time-to-flare. Applying the methodology described above to these data yielded averaged hazard ratio of 0.60 (95% CI: 0.28-0.91, width = 0.63, p = 0.02) for sex and 1.98 (95% CI: 0.83-3.12, width = 2.29, p = 0.01) for methotrexate use, indicating both are significantly associated with flare occurrence, with CI widths on the log-hazard ratio scale reduced by 5.54% and 36%, respectively, compared with Rayner et al.[2].

Discussion

The proposed method improves efficiency over standard approaches for continuous variables as endpoints, demonstrated through simulations and application to the Bio-Flare study.

[1]10.1016/j.jspi.2020.02.003

[2]10.1136/bmjopen-2024-092478

Double inverse probability weighting for modelling sparse longitudinal markers in sub-cohort with left-and-right truncation by death

12:15-12:30

Presenter: Trinh Dong

Authors: Trinh Dong [1, 2]; Hélène Jacqmin-Gadda [3]; Denis Rustand [4]; Alexandra Samier Foubert [2]; Cécile Proust-Lima [5]

Affiliations: [1] Population Health Sciences Institute, Newcastle University, UK, Newcastle upon Tyne; [2] Univ. Bordeaux, INSERM, Institut Bergonié, BPH, U1219, CIC 1401, Bordeaux; [3] National Institute of Health and Medical Research (Inserm U1219, Bordeaux population health), University of Bordeaux, Bordeaux; [4] Inserm U1219, University of Bordeaux, Bordeaux; [5] Université de Bordeaux, Inserm, Bordeaux Population Health Research Center, U1219, Bordeaux

Introduction

In epidemiological research, it is common for cohort participants to be recruited into a sub-cohort after initial enrolment. For instance, in the French Multiple System Atrophy (F-MSA) cohort where patients are followed annually from diagnosis, a sub-cohort was initiated to study in-depth neuro-imaging biomarkers. For high-mortality conditions like MSA, delayed recruitment induces left-truncation bias: individuals who died before sub-cohort creation are systematically excluded and under-represented. Furthermore, subsequent deaths post-recruitment can lead to right truncation. In the context of a sparsely measured outcome, the reliability of joint modelling is limited.

To address this, we propose a double inverse probability weighting (dIPW) method for mixed effects models (LMMs) that corrects both left- and right-truncation bias by leveraging information from a second densely measured biomarker that is available in the full cohort and mediates the pathway between the primary outcome and survival

Methods

The dIPW technique is built upon the dynamic prediction of participant survival at entry in the sub-cohort and at subsequent visits. In brief, the survival probability of each sub-cohort observation is estimated from a joint model for the longitudinal mediator and survival processes fitted on the full cohort. Inverse probabilities are then applied on the modified mixed-effects model at two levels: (1) subject-specific weight for the individual random effects and (2) visit-specific weight per each outcome measurement. Bayesian estimation is done using Hamiltonian Monte Carlo via a modified brms/Stan model, with uncertainty quantified through parametric bootstrap.

Results

We assessed the performance of the proposed dIPW-LMM in simulations. Under positivity assumption, the dIPW-LMM showed minimal bias and correct 95% credible intervals coverage while the unweighted LMM neglecting left-and-right-truncation showed substantial bias, especially for time-related effects.

The method was applied in the neuro-imaging sub-cohort (N=176) of F-MSA study (N=783) to assess the differences in brain atrophies according to Cerebellar and Parkinsonian MSA subtypes.

Discussion

By leveraging a mediator densely measured on the full cohort, dIPW-LMM reduces death-related truncation bias in sub-cohort analyses. Even without left truncation, the proposed method offers an alternative to joint modelling to reduce attrition bias, particularly when complex and/or high-dimensional biomarkers are sparsely collected. Study limitations include simple data generating process, positivity assumption, and the fully mediated pathway assumption between the primary outcome and survival.

Small Sample and Sequential Trial Methods

Tue, 29.09.26, 11:00–12:30, Room: K5

Chair: Richard Tawiah

Stratified Analysis of Small-Sample Clinical Trials with Binary Endpoints

11:00-11:15

Presenter: Vanessa Ihl

Authors: Vanessa Ihl [1]; Ralf-Dieter Hilgers [2]

Affiliations: [1] RWTH Aachen University, Aachen; [2] Sigmund Freud Private University Vienna, Vienna

Introduction

Small sample sizes are common in clinical trials for rare diseases or early-phase studies. When the outcomes are binary, standard large-sample methods, such as the chi-squared test, may perform poorly due to low power and unreliable asymptotic assumptions. Stratification by prognostic factors (e.g., age, study centre, or disease severity) is used and encouraged (e.g., in the ICH E9 guideline) to improve comparability and reduce confounding between treatment groups. We introduce a novel exact stratified test, based on Boschloo's test, and demonstrate the advantages of using appropriate testing strategies, using a clinical trial as showcase.

Methods

We consider parallel group, stratified randomized clinical trials with small sample sizes and a binary primary endpoint. Analytical approaches include exact methods such as the (stratified) Fisher's exact test and Boschloo's unconditional exact test. Based on these we build a stratified version of Boschloo's test and compare it with the aforementioned tests and the Cochran-Mantel-Haenszel test. We present a showcase of a clinical trial conducted for a rare disease to demonstrate the importance of applying the appropriate testing strategy and show results of clinical trial simulations we conducted across different randomization procedures to evaluate type I error control and power for several trial designs.

Results

Standard exact methods are rather underpowered, since they rely on larger sample sizes. Existing exact tests like the stratified Fisher's test control type I error but are conservative, particularly with sparse data. The stratified version of Boschloo's unconditional test is less conservative and reaches higher power levels while maintaining the type I error. This leads to more efficient designs and reduced sample size requirements, as we demonstrate in accompanying sample size calculations.

Discussion

In small-sample trials with binary endpoints, particularly when stratification is employed, standard methods may result in inefficient inference. The proposed stratified unconditional exact tests improve efficiency without compromising validity.

These gains are particularly relevant in rare disease settings, where minimizing sample size is critical. Implementation may require greater computational effort, but improved software availability could facilitate broader use.

Small sample-corrected variance estimators for covariate adjustment using weighting in individually randomized trials

11:15-11:30

Presenter: Elizabeth Williamson

Authors: Elizabeth Williamson [1]; Suzie Cro [2]; Tim Morris [3]

Affiliations: [1] London School of Hygiene and Tropical Medicine, London; [2] Imperial College, London; [3] Novartis, London

Introduction

Covariate adjustment is becoming increasingly popular in the analysis of randomised trials, in order to enhance precision in estimating the treatment effect. One way of doing this is by using weighting, for example inverse probability of treatment weighting (IPTW). Despite several desirable statistical properties, the standard sandwich variance estimator means inference can be overly liberal at small sample sizes.

Methods

Following similar approaches used in the context of cluster randomised trials, we explore a range of corrections, based on the leverage, to the sandwich variance estimator, along with a correction based on the degrees of freedom. We perform a simulation study, using sample sizes from 40 to 200, comparing this correction to an alternative leverage-based correction, a simple degrees-of-freedom correction and the uncorrected estimator. We also explore the impact of adopting a t-distribution in place of the usual Z-distribution for calculation of confidence intervals and p-values. We apply these methods to two types of weight: ATE (average treatment effect) weights and overlap weights. We apply these methods to re-analyse the ADAPT trial, which included 60 children with severe eczema.

Results

The uncorrected variance estimator resulted in considerable under-coverage at smaller sample sizes, but less so with increasing correlation between covariates. The degrees-of-freedom correction performed well with independent covariates but over-covered with correlated covariates. The leverage-based corrections showed variable performance under different settings.

Discussion

Small-sample corrections to the variance of IPTW estimators, using either a simple degrees-of-freedom estimator or our proposed leverage-based correction, result in more accurate inference for weighted estimators, provided that there is little correlation among the covariates being adjusted for. These corrections should be considered in randomised trials with fewer than 200 participants to ensure valid statistical inference when using IPTW, unless substantial correlations are expected among covariates. Similarly, the t-distribution should be used for inference with continuous outcomes.

Unbiased Estimation in Adaptive Enrichment Designs

11:30-11:45

Presenter: Enyu Li

Authors: Enyu Li [1]; Nigel Stallard [2]; Ekkehard Glimm [3]; Dominic Magirr [3]; Kimani Peter [2]

Affiliations: [1] University of Warwick, Coventry; [2] Warwick Clinical Trials Unit, University of Warwick, Coventry; [3] Novartis Pharma AG, Basel

Introduction

Recent advances in biomedical research have identified numerous biomarkers that may be related to the success of medical treatments. When a treatment is suspected to benefit only certain patient subgroups, it can be more efficient and ethical to use adaptive enrichment designs, where a pre-planned interim analysis is incorporated to select a promising patient subgroup based on accrued data. However, these data-dependent adaptations introduce significant statistical challenges. If the conventional maximum likelihood estimator (MLE) is used, the treatment effect in the selected subpopulation is often overestimated. Currently, inference methods to correct this bias are typically tailored to specific interim decision rules, meaning an unbiased estimator must be re-derived if a new decision rule is employed. Hence, there remains a pressing need for estimation frameworks that can encompass a broader class of subpopulation selection rules.

Methods

We define a general class of interim decision rules based on a sample space partition condition. Based on the Rao-Blackwell Theorem and the Lehmann-Scheffé Theorem, we provide a systematic derivation that yields a unified formula for the Uniformly Minimum Variance Conditional Unbiased Estimator (UMVCUE) for any adaptive enrichment design with an interim decision rule in this class. This adaptability allows our formulation to encompass a wide spectrum of enrichment designs, including several existing rule-specific UMVCUEs as special cases, thereby confirming the generality of the proposed framework.

Results

Through extensive simulations, we evaluate the proposed UMVCUE against the MLE. In all considered scenarios, the bias of the UMVCUE is negligible, confirming its conditional unbiasedness. In contrast, the MLE exhibits substantial bias, particularly when a mismatched subpopulation is selected. These results highlight that the UMVCUE ensures therapeutic benefits are not overestimated.

Discussion

We have established a unified theoretical framework that extends conditional unbiased estimation to a broad class of adaptive enrichment designs, effectively bridging the gap between flexible interim subpopulation selection and rigorous statistical inference. By providing a systematic approach to deriving the UMVCUE, our work eliminates the necessity for case-specific derivations for each new design. This facilitates the implementation of diverse and flexible interim rules with greater ease in real-world trials.

Anytime-valid testing with e-values and confirmatory adaptive designs

11:45-12:00

Presenter: Werner Brannath

Authors: Werner Brannath; Lasse Fischer

Affiliations: University of Bremen, Bremen

Introduction

Confirmatory adaptive designs were introduced more than 30 years ago and enable for example sample size re-assessments and the selection of treatments, endpoints as well as subpopulations during the course of a clinical trial. Recently, sequential tests based on e-values for an anytime-valid inference have been developed, promising seemingly similar or even more flexibility and utility. In this talk, we formally compare these two independently developed concepts.

Methods

We introduce e-value based, anytime-valid sequential tests and present a unified approach for confirmatory adaptive designs based on a sequence of conditional error functions (implicitly introduced in earlier work). We show mathematically that these two approaches are formally equivalent. We also indicate how non-exhaustive e-value based sequential tests may be exhausted (and thereby uniformly improved) by recursive combination tests. In addition, we review confirmatory adaptive multiple testing with partial conditional error functions and link this approach to an equivalent e-value combination approach which is the only one that is admissible under general distributional assumptions. We finally illustrate how e-values can be used to introduce flexibility with regard to the level and loss function to be controlled in a confirmatory trial.

Results

In spite of their common formal equivalence, confirmatory adaptive designs and e-value based sequential tests have quite different emphases: While hypothesis testing with combination tests and conditional error function usually intent to exhaust type I error rates under the offered flexibility, e-value based testing aims on the additional flexibility with regard to distributional assumptions, optional continuation, the chosen level and, in recent extensions, in a general and flexible control of loss functions.

Discussion

This work sheds light on the formal and methodological connections and differences between confirmatory adaptive designs and e-value based, anytime-valid sequential tests.

Bayesian sequential monitoring and analysis of randomised controlled trials

12:00-12:15

Presenter: Luana Boumendil

Authors: Luana Boumendil; Haiyan Zheng

Affiliations: Department of Mathematical Sciences, University of Bath, Bath

Introduction

Group sequential designs enable early stopping, for efficacy or futility, reducing clinical trial costs and duration while accelerating access to effective treatments and limiting exposure to ineffective ones. Decision boundaries are typically defined by controlling frequentist error rates. Bayesian designs, allowing incorporation of external information and direct modelling of uncertainty, are increasingly adopted.

Methods

We propose an easy-to-implement approach to Bayesian sequential monitoring of K-stage randomised controlled trials, for which the stage-wise error rates are controlled at a desired level while allowing early stopping for efficacy or futility. Similarly to classical group sequential designs with a frequentist analysis, continuation and stopping probabilities can be computed either via conditional probabilities or via joint multivariate normal modelling (Jennison & Turnbull, 1999). Decision rules based on the posterior probability of efficacy can be equivalently expressed as a comparison between the observation at stage k and stopping boundaries that are conditional on accumulated observations from previous stages 1, ...,k-1 (Freedman & Spiegelhalter, 1989). Under this formulation, stage-wise frequentist error rates of this Bayesian design can be obtained as the conditional probabilities of crossing these boundaries, given that no stopping has occurred at earlier analyses. However, the conditional probability formulation requires the computation of nested integrals, which become infeasible as the number of analyses increases. By jointly modelling the decision rules as a multivariate

normal distribution, the Bayesian stopping rules can be expressed as linear constraints through a linear transformation of the Bayesian stopping probabilities, exploiting the symmetry property of the Normal distribution.

Results

One feature of the proposed approach to sequential monitoring is that prior information can be incorporated to determine the decision boundaries. The correctness of the derivations was verified through simulation.

Discussion

Finally, the applicability of the method to a time-to-event endpoint under an exponential model, with independent log-normal priors on the median survival in each arm, is illustrated using the COBRA trial (NIHR160459).

Jennison, C., & Turnbull, B. W. (1999). Group sequential methods with applications to clinical trials. CRC Press. Freedman, L. S., & Spiegelhalter, D. J. (1989). Comparison of Bayesian with group sequential methods for monitoring clinical trials. Controlled clinical trials, 357–367.

Bootstrap Prediction Intervals for Interim Event Counts in Time-to-Event Clinical Trials

12:15-12:30

Presenter: Edoardo Ratti, Federico Luigi Perlino

Authors: Edoardo Ratti [1]; Federico Luigi Perlino [2]; Stefania Galimberti [1]; Maria Grazia Valsecchi [1]

Affiliations: [1] Bicocca Center of Bioinformatics, Biostatistics and Bioimaging (B4 centre), School of Medicine and Surgery, University of Milano-Bicocca, Monza; [2] Department of Statistics, University of Warwick, Coventry

Introduction

Time-to-event endpoints are widely used in clinical trials. In group-sequential designs, interim and final analyses are driven by the amount of accumulated information, which typically depends on the number of observed events. Although survival assumptions are set at the design stage to achieve the target level of power, unexpected changes in event rates may affect both planned operating characteristics and trial logistics. Predicting the number of events expected at planned future looks, along with the uncertainty around those predictions, is therefore important.

Methods

Rather than predicting the time needed to reach a target number of events, we focus on predicting the number of events by a fixed future date with prediction intervals. Motivated by a phase III trial in pediatric acute lymphoblastic leukaemia (ALL), we extend to interim monitoring a framework originally developed in reliability engineering1. The extended method accommodates random drop-out,

staggered entry dates, patient-level covariates, and possible dependence between entry dates and loss to follow-up. Prediction intervals are obtained by bootstrap approximation of the predictive distribution of future events, keeping event and censoring indicators fixed and resampling event times from right-truncated fitted parametric models.

Results

In the pediatric ALL application, monthly predictions were broadly consistent with the observed event counts, but differed across parametric models, especially for longer extrapolations, highlighting model sensitivity. Royston-Parmar models helped capture departures from simpler parametric assumptions. In simulations, bootstrap prediction intervals achieved near-nominal coverage overall, with worse performance for longer horizons and less mature follow-up. Loss to follow-up had only a modest impact. Flexible models were not consistently superior, and over-parameterization could reduce coverage.

Discussion

This work provides a novel framework for interval prediction of the number of events at future analyses in clinical trials with planned interim looks. Greater model flexibility does not automatically improve prediction performance. In practice, follow-up maturity and prediction horizon are the main determinants of interval accuracy, so model choice should be guided by extrapolative plausibility and sensitivity analyses.

1Tian Q., Meng F., Nordman D. J., Meeker W. Q., Predicting the Number of Future Events. Journal of the American Statistical Association. 2021 117(539): 1296-1310

Robust Causal and Statistical Methods for Complex Data

Tue, 29.09.26, 11:00–12:30, Room: N1

Chair: Paola Rebora

Assessing covariate balance after weighting in propensity score analysis using weighted z-difference

11:00-11:15

Presenter: Chengshan Ma

Authors: Chengshan Ma [1]; Andreas Wienke [2]; Oliver Kuß [3]; Alexandra David [1]

Affiliations: [1] Institute of Medical Epidemiology, Biometry and Informatics of Martin-Luther-University Halle-Wittenberg, Halle (Saale); [2] Martin-Luther University Halle-Wittenberg, Institute of Medical Epidemiology, Biostatistics and Informatics, Halle (Saale); [3] Deutsches Diabetes-Zentrum (DDZ), Leibniz-Zentrum für Diabetes-Forschung an der Heinrich-Heine-Universität Düsseldorf, Düsseldorf

Introduction

In observational studies, propensity score (PS) analysis is widely used to reduce confounding by balancing baseline covariates. The commonly used standardized difference for assessing balance has limitations: its distribution depends on sample size and it only applies to continuous and binary variables. The measure z-difference (Kuss, 2013) overcomes these issues. However, it is only applicable in unweighted settings. This study extends the z-difference to weighted settings, enabling imbalance diagnosis after PS weighting.

Methods

Weighted z-differences were derived for continuous, binary and nominal variables based on existing formulas, with sample estimates replaced by their weighted counterparts. For ordinal variable, the weighted rank mean difference was used with variance obtained via bootstrapping method. A simulation study was conducted under two scenarios: no baseline imbalance and the presence of baseline imbalance. Two weighting methods were applied - inverse probability of treatment weighting (IPTW) targeting the average treatment effect on the entire population (ATE) and overlap weighting targeting the average treatment effect in the overlap population (ATO). Performance of weighted z difference was further evaluated by examining its bias and mean squared error (MSE). This measure was also applied to real-world data from a published PS analysis (Börgermann et.al. 2012).

Results

Across simulation scenarios, all covariates achieve good balance, the weighted z-differences were generally close to zero, with mean absolute values below the rule-of-thumb threshold 1.96 in most settings. The results showed small bias under IPTW, though MSE was slightly elevated due to weight variability. Application to real-world data confirmed these findings.

Discussion

The weighted z-difference appears to be a useful measure to assess covariate balance after weighting in PS analysis. It addresses key limitations of the standardized difference, offering sample size-independent inference and applicability to all variable types.

Reference:

Kuss O. (2013). The z-difference can be used to measure covariate balance in matched propensity score analyses. Journal of Clinical Epidemiology, 66(11), 1302-1307.

Börgermann J., et.al. (2012). Clampless off-pump versus conventional coronary artery revascularization: a propensity score analysis of 788 patients. Circulation 126, S176-182.

Effects of Time Trends on Risk Estimation in Self-Controlled Designs - A Simulation Study

11:15-11:30

Presenter: Katharina Meiszl

Authors: Katharina Meiszl [1]; Sara Schramm [2, 3]; Nina Timmesfeld [1]

Affiliations: [1] Ruhr University Bochum, Department of Medical Informatics, Biometry and Epidemiology, Bochum; [2] Fliedner University of Applied Sciences Düsseldorf, Düsseldorf; [3] Institute for Medical Informatics, Biometry and Epidemiology, University Hospital Essen, University Duisburg-Essen, Essen

Introduction

For examining the relationship between exposure and outcome, self-controlled designs (SCDs) provide a robust approach by adjusting for time-independent confounders within individuals, but are sensitive to temporal changes, including time-dependent confounders, seasonality, and time trends [1]. This detailed simulation study aims to assess the susceptibility of SCDs to different types of time trends.

Methods

Discrete time simulation was used for data generation [2], and the analysis was performed using Monte Carlo simulation [3]. The study considered various time trends in exposure and outcome. Risk modification was described as relative risk. Three SCDs were compared: the self-controlled case series [4], the case-crossover design [5], and the sequence symmetry analysis [6]. The results were evaluated using quality measures assessing the impact of time trends on the estimators.

Results

Our simulation study indicated that all three SCDs were susceptible to bias in the presence of time trends. The extent of bias varying depending on the specific design and the shape of the time trend.

Discussion

When using SCDs for extended observation periods, it is essential to first examine and comprehensively describe the data in terms of potential time trends before conducting further analyses. Our findings highlight the importance of selecting an analysis method that aligns with the characteristics of the data and potential temporal changes.

[1] Takeuchi Y, Shinozaki T, Matsuyama Y. A comparison of estimators from self-controlled case series, case-crossover design, and sequence symmetry analysis for pharmacoepidemiological studies. BMC Medical Research Methodology. 201818(1). doi:10.1186/s12874-017-0457-7.

[2] Tang J, Leu G, Abbass HA. Discrete Time Simulation. Simulation and Computational Red Teaming for Problem Solving. 2019143-55. doi:10.1002/9781119527183.ch8.

[3] Mooney C. Monte Carlo Simulation. SAGE Publications, Inc. 1997. doi:10.4135/9781412985116

[4] Farrington P, Rush M, Miller E, Pugh S, Colville A, Flower A, et al. A new method for active surveillance of adverse events from diphtheria/tetanus/pertussis and measles/mumps/rubella vaccines. The Lancet. 1995345(8949):567-9. doi:10.1016/s0140-6736(95)90471-9.

[5] Maclure M. The Case-Crossover Design: A Method for Studying Transient Effects on the Risk of Acute Events. American Journal of Epidemiology. 1991133(2):144-53. doi:10.1093/oxfordjournals.aje.a115853.

[6] Hallas J. Evidence of Depression Provoked by Cardiovascular Medication: A Prescription Sequence Symmetry Analysis. Epidemiology. 19967(5):478-84. doi:10.1097/00001648-199609000-00005.

Quantitative Bias Analysis as Distributionally Robust Optimization

11:30-11:45

Presenter: Shu Tamano

Authors: Shu Tamano [1, 2]; Yui Tomo [2]

Affiliations: [1] The University of Tokyo, Tokyo; [2] Japan Institute for Health Security, Tokyo

Introduction

Quantitative bias analysis (QBA) provides essential tools for assessing systematic errors in epidemiology, yet it remains methodologically fragmented. Current practices rely on disparate, bias-specific parameters such as E-values for confounding or misclassification rates, which complicates joint evaluations. Because QBA fundamentally bounds causal estimates under worst-case residual bias, it naturally aligns with distributionally robust optimization (DRO). However, applying standard DRO directly to observed data inappropriately conflates latent biases with sampling variation and model misspecification.

Methods

We propose a unified QBA framework built on structural Wasserstein ambiguity sets. Rather than perturbing the observed-data distribution, our approach holds the observed margin strictly fixed and exclusively perturbs the conditional law of latent, bias-relevant variables (e.g., potential outcomes or unobserved true covariates). The magnitude of this perturbation is constrained by a structural Wasserstein radius rho. By tailoring the ground metric on the latent space, for instance using a weighted Hamming cost for discrete misclassification or absolute deviation for continuous missingness, a single scalar budget rho transparently governs single or simultaneous hidden biases.

Results

We establish a conditional dual representation that characterizes the sharp upper and lower sensitivity envelopes for generic mean-type estimands. For continuous and bounded binary causal functionals, we prove that these DRO envelopes admit exact, closed-form algebraic solutions under mild transport-richness conditions. For continuous outcomes, the worst-case bias bounds scale exactly linearly, while bounded binary targets are explicitly clipped by their natural parameter space. We leverage these theoretical guarantees to wrap any asymptotically linear benchmark estimator with distributionally robust confidence intervals and rigorous one-sided tests. Additionally, we introduce the transport robustness value, which quantifies the minimal latent structural edit budget required to overturn an epidemiologic finding.

Discussion

This framework subsumes and translates existing deterministic and probabilistic QBA methods into a cohesive DRO problem. By shifting the sensitivity parameter to a structurally interpretable transport metric, epidemiologists can systematically evaluate and communicate the fragility of studies to complex, interacting biases using a mathematically unified, computationally tractable methodology.

Simulation study on feature selection methods for longitudinal, high-dimensional biomedical data

11:45-12:00

Presenter: Alexander Gieswinkel

Authors: Alexander Gieswinkel [1, 2, 3]; Gökhan Gül [1, 3, 4]; Gregor Buch [1, 3]; Vincent ten Cate [1, 4, 3]; Lisa Hartung [2]; Philipp S. Wild [1, 4, 3, 5]

Affiliations: [1] Preventive Cardiology and Preventive Medicine, Department of Cardiology, University Medical Center of the Johannes Gutenberg University Mainz, Mainz; [2] Institute of Mathematics, Johannes Gutenberg University Mainz, Mainz; [3] German Center for Cardiovascular Research (DZHK), partner site Rhine Main, Mainz; [4] Clinical Epidemiology and Systems Medicine, Center for Thrombosis and Hemostasis (CTH), University Medical Center of the Johannes Gutenberg University Mainz, Mainz; [5] Institute of Molecular Biology (IMB), Mainz

Introduction

Longitudinal observations on biomedical (e.g., omics) data are beneficial to characterize the progression or onset of diseases with greater reliability than cross-sectional data. Due to advances in high-throughput technologies, prospective cohort studies equipped with high-dimensional (p gg N) data are becoming increasingly common. This circumstance requests supervised feature selection methods tailored to longitudinal data. Our prior systematic review of available R packages on CRAN (Comprehensive R Archive Network) revealed 13 suitable methods. To systematically compare these methods in realistic scenarios, a simulation study was performed.

Methods

The simulation study was conducted according to the ADEMP framework. Interclass and intraclass correlation structure of the simulated data was generated based on the structure of longitudinal and high-dimensional proteomics data from a prospective cohort study. Two scenarios sets with different number of features were compared: Scenario set 1 captured p=100 features (true predictors pt=20) with Ninlbrace100,500,1000rbrace samples and scenario set 2 considered p=1000 features (pt=100) with sample sizes Ninlbrace100,500,1000,5000rbrace. All scenarios were simulated using tinlbrace3,5,8rbrace timepoint, resulting in 21 total settings. For evaluation, the true positive rate (TPR) and false discovery rate (FDR) were merged to a performance score PS:=frac12TPR+frac12(1-FDR).

Results

In scenario set 1, the TPR increased by 3.7 %pt. (percentage point) per 100 additional samples and by 3.9 %pt. per additional timepoint, while the FDR decreases by 0.6 %pt. with increasing N and by 0.7 %pt. with increasing t. Scenario set 2 revealed similar but weaker effects. The Bayesian spikeSlabGAM outperformed any other method in scenario set 1 at N=1000, regardless of time points (e.g. PS=88% at t=3). Aside from the frequentist sgee package, it was the only method converging in scenario set 2 at N=100. However, both methods performed equally poorly in this scenario e.g. at t=8 (TPR=4%, FDR=70%). At N=5000 and t=8, another Bayesian method, sparsereg, performed best (PS=64%), followed by sgee (PS=61%), while spikeSlabGAM did not converge at this dimension and returned error messages in each simulation run.

Discussion

Repeated measurements strengthen the reliability on sparse model building. In the observed p≤N scenarios, of the existing approaches, the Bayesian spikeSlabGAM and sparsereg tended to perform best. Feature selection in p gg N scenarios remains challenging, even when considering longitudinal data. More work is needed to improve the performance in such scenarios.

Foundation Models and Causal Inference - A Paradigm Shift for Precision Medicine?

12:00-12:15

Presenter: Karl Köchert

Authors: Karl Köchert [1]; Nils Ternès [2]; Hugo Hadjur [3]; Laura Schlieker [4]; Adam Skubala [5]; Eliana Garcia Cossio [6]; Antigoni Elefsinioti [6]; Maike Ahrens [7]; Ruben Sethi [8]; Sebastian Voß [7]

Affiliations: [1] Sanofi-Aventis Deutschland GmbH, Berlin; [2] Sanofi, Montpellier; [3] Saryga, Paris; [4] Staburo GmbH, Munich; [5] Bayer AG, Wuppertal; [6] Bayer AG, Berlin; [7] Evidenze, Essen; [8] Stratum Bio, Paris

Introduction

Quantitative and qualitative evaluation of prognostic and predictive biomarkers is central to decision-making in precision medicine drug development. Advances in high-throughput technologies now enable simultaneous measurement of thousands to millions of biomarkers across molecular layers, creating both opportunity and analytical challenges. Traditional statistical approaches struggle with the scale, complexity, and nonlinear structure of such data. This has accelerated interest in combining causal inference frameworks with modern artificial intelligence (AI) and machine learning (ML) methods. In particular, pretrained foundation models for tabular data, such as TabPFN [1], hold promise for transforming how complex multimodal datasets are analyzed, facilitating the selection of optimal therapeutic interventions.

Methods

To evaluate a causal inference framework that leverages foundation models as base learners, the ML subgroup of the PSI/EFPSI Special Interest Group on Biomarkers launched a hands-on project focused on treatment effect heterogeneity in high-dimensional biological settings. The working model organizes contributors into two agile teams: a simulation team and a methods benchmarking team.

The simulation team generates synthetic biomarker datasets with increasing complexity across project cycles, using both classical and AI-assisted approaches. Key parameters include p-to-n ratios, monotonicity and nonlinearity of predictor-outcome relationships, higher order interactions, and correlation structures reflective of real gene expression networks.

The methods team, blinded to the data generation process, analyzes the datasets using causal inference meta-learners built on classical ML techniques and the TabPFN model. Their objective is to recover the underlying data generating process, identify true prognostic and predictive biomarkers, and estimate conditional average treatment effects (CATE).

Results

We will present benchmarking results comparing TabPFN with established ML methods across data scenarios of varying dimensionality and complexity. Performance will be evaluated in terms of correct biomarker identification, CATE estimation, and robustness to challenging data structures.

Discussion

Our findings illustrate the promise of foundation models such as TabPFN for high-dimensional precision medicine applications. We will also highlight how the interactive dual team working model fostered engagement, creativity, and methodological rigor, offering a scalable approach for future collaborative ML-driven biomarker research.

References

1. Hollmann et al. Accurate predictions on small data with a tabular foundation model. Nature (2025).

Identifying harmful drug-drug interactions using Bayesian Profile Regression and stochastic variational inference applied to population-scale linked routine health data in Wales

12:15-12:30

This is one of this year's ISCB StCA award-winning abstracts.

Presenter: Timothy Osborne

Authors: Timothy Osborne; James Rafferty; Rhiannon K Owen

Affiliations: Swansea University, Swansea

Introduction

Polypharmacy is common among people with multiple long-term conditions (MLTC) and can increase the risk of medication-related harm (MRH). Most approaches focus on pairwise drug interactions, and do not scale well to higher-order interactions in patients taking many medications. We aimed to identify clusters of prescribed medications associated with MRH using linked routine health data.

Methods

We used linked primary care, hospital, and demographic data within the Secure Anonymised Information Linkage (SAIL) Databank to construct a retrospective cohort of all adults living in Wales with MLTC in 2019. Medication profiles were defined using Anatomical Therapeutic Chemical (ATC) level 3 classes, using prescription records from primary care. MRH was defined as a hospital-recorded adverse drug event within 30 days of a new prescription, identified using ICD-10 codes. Bayesian Profile Regression (BPR) using full-rank stochastic variational inference (SVI) was applied to cluster medication profiles while conditioning on MRH and adjusting for age, sex, and deprivation.

Results

In the 55-74 age group (n=553,641), the model identified 17 distinct clusters. Most individuals belonged to larger, lower-risk clusters characterised by routine prescribing profiles. The highest-risk cluster (MRH probability 0.22) was characterised by higher probabilities of drugs for constipation, opioids, and antihistamines. The second highest-risk cluster (MRH probability 0.08) showed a complex diabetes-related profile, with insulins and analogues, other blood glucose-lowering drugs, antithrombotic agents, and drugs for peptic ulcer and GORD. Comparisons between similar clusters with differing probabilities of MRH revealed which combinations may be driving risk. Two clusters with heavy cardiovascular prescribing resulted in different risk of MRH the higher-risk cluster (MRH probability 0.05) had additional respiratory and pain medications and a generally more complex prescribing profile than the lower risk cluster (MRH probability 0.01). Across clusters, drugs for peptic ulcer and GORD, opioids, analgesics, penicillin, and antidepressants appeared more frequently in higher-risk profiles.

Discussion

BPR using SVI provides a scalable approach for modelling higher-order medication combinations in population-level prescribing data. By identifying clinically interpretable clusters associated with MRH, it can help to inform policy on potentially harmful medication combinations in people living with MLTC.

Further work has explored more granular ATC classifications, and additional age groups.

Data Foundations for Clinical AI

Tue, 29.09.26, 11:00–12:30, Room: N2

Chair: Stefan Lenz

Refining and Scaling the GeMTeX Semantic Annotation for German Clinical Text

11:00-11:15

Presenter: Justin Hofenbitzer

Authors: Justin Hofenbitzer [1]; Christina Lohr [2]; Luise Modersohn [3]; Frank A. Meineke [4]; Markus Löffler [5]; Martin Boeker [6]

Affiliations: [1] Technical University of Munich, TUM School of Medicine and Health, Institute for AI and Informatics in Medicine (AIIM), Chair of Medical Informatics, TUM University Hospital, Munich; [2] IMISE - Uni Leipzig, Leipzig; [3] Institute for Artificial Intelligence and Informatics in Medicine, Chair of Medical Informatics, Medical Center rechts der Isar, Technical University of Munich, Munich, Germany, München; [4] Universität Leipzig, Leipzig; [5] Universität Leipzig, Med. Fakultät, Leipzig; [6] Technische Universität München (TUM), München

Introduction

Clinical free text contains substantial medically relevant information often absent from structured data [1]. However, semantically enriched German clinical corpora remain scarce, limiting interoperability and developments of AI applications [2,3]. The multicentric German Medical Text Corpus (GeMTeX) addresses this gap through semantic annotation of clinical documents using SNOMED CT [4]. This contribution reports recent progress in operationalizing and refining the GeMTeX semantic annotation campaign.

Methods

All annotators are medical students who have completed their second year and were trained according to a comprehensive annotation guideline. After pre-annotation, all documents are manually annotated using a recommender system. To monitor reliability, 1% of all documents per site are annotated by multiple annotators to estimate agreement. Bi-weekly adjudication rounds with curators and meetings within the annotation groups are conducted to review difficult cases. The workflow is designed to balance annotation progress with methodological rigor.

Results

As of March 2026, GeMTeX comprises 503 semantically annotated documents with 220k annotations. The terminology-constrained annotation campaign ensures quality control and iterative adjudication. A synthetic gold-standard document is available as a methodological resource (https://doi.org/10.5281/odo.18861607).

Discussion

GeMTeX advances the creation of a semantically interoperable German clinical text resource usable via the FDPG. The project demonstrates substantial progress in annotation, with a detailed, quality-assured workflow.

References

[1] Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review. JMIR Med Inform. 2019 Apr 277(2):e12239. doi:10.2196/12239

[2] Névéol A, Dalianis H, Velupillai S, Savova G, Zweigenbaum P. Clinical Natural Language Processing in languages other than English: opportunities and challenges. J Biomed Semant. 2018 Dec9(1):12. doi:10.1186/s13326-018-0179-8

[3] Hahn U. Clinical document corpora - real ones, translated and synthetic substitutes, and assorted domain proxies: a survey of diversity in corpus design, with focus on German text data. JAMIA Open. 2025 May 28(3):ooaf024. doi:10.1093/jamiaopen/ooaf024

[4] Hofenbitzer J, Schulz S, Boeker M, Klügl P, Riepenhausen S, Lohr C, et al. Introducing Medical Semantic Annotation Guidelines for German Clinical Documentation with SNOMED CT. 70 Jahrestag Dtsch Ges Für Med Inform Biom Epidemiol E V GMDS. 2025. doi:10.3205/25GMDS108

Automated Extraction of Genetic Eligibility Criteria from Clinical Trial Records Using LLMs - A Technical Case Report

11:15-11:30

Presenter: Georg Mathes

Authors: Georg Mathes [1]; Stephanie Berger [2]; Stefan Sigle [1]

Affiliations: [1] MOLIT Institute gGmbH, Heilbronn; [2] SLK Kliniken Heilbronn, Heilbronn

Introduction

Accurate interpretation of clinical trial eligibility criteria is essential for applications such as patient-trial matching and clinical decision support, particularly in precision oncology. However, relevant information, including genetic mutation requirements, is typically embedded in unstructured text within trial registries such as ClinicalTrials.gov, limiting its accessibility for automated processing.

Methods

This paper presents a technical case report on the development and integration of a system for extracting and structuring mutational eligibility criteria from clinical trial records. The system focuses on identifying mutated genes, distinguishing inclusion and exclusion criteria, and assigning

them to individual study arms. To address challenges such as ambiguous abbreviations and context-dependent meaning, the approach combines large language models (LLMs) for context-aware extraction with rule-based validation against HGNC gene nomenclature. The system was implemented using locally deployed LLMs to ensure data privacy and reproducibility and integrated into the ETL pipeline of the Community Annotated Trial Search (CATS) platform.

Results

Applied to 4,918 clinical trials, the system generated structured representations of genetic eligibility criteria for 1,010 studies. Evaluation based on expert review of 42 trials showed that 88.1% of studies were correctly annotated. At the level of individual eligibility criteria, a precision of 80% was achieved. Remaining errors were mainly due to hallucinated gene mentions and misinterpretation of abbreviations, highlighting challenges in biomedical text processing.

Discussion

The findings indicate that LLM-based extraction is a promising approach for structuring complex eligibility criteria, particularly when combined with validation strategies to improve precision. The integration into an operational system demonstrates practical feasibility, while the observed limitations emphasize the need for careful dataset design, error mitigation strategies, and continued refinement to achieve reliable automation in clinical applications.

The full paper of this abstract will be published in the GMDS series "Studies in Health Technology and Informatics" with doi: 10.3233/SHTI260998, URL:https://dx.doi.org/10.3233/SHTI260998

The Estimand Begins in the Data: How Dataset Construction Shapes Clinical Trial Inference

11:30-11:45

Presenter/Author: Nicole Close, Vasyl Kornyushyn, Yaroslav Sud

Affiliations: EmpiriStat, Inc., International

Introduction

Clinical trial inference is often assumed to begin with statistical modelling of a predefined analysis dataset. In practice, however, the inferential target is frequently shaped earlier, through decisions about how raw data are curated, reconciled, and transformed into analysable structures. Choices made during data management and dataset construction determine which observations are retained, how time is represented, and how outcomes are derived. We examine how this upstream data architecture influences the estimand that is ultimately realised in analysis.

Methods

We reviewed recurring data-construction challenges observed across recent industry clinical programmes in neurodegenerative disease, anti-infective development, and metabolic trials. Examples

included: reconstruction of longitudinal cognitive outcomes when visits were partially completed or performed outside protocol windows adjudication of microbiological response timing when cultures and clinical assessments occurred on different days and definition of baseline and exposure in titration-based metabolic studies with dose interruptions or restarts. For each situation, alternative operationally plausible dataset construction strategies were applied and identical statistical analyses were then performed to assess downstream effects on treatment interpretation.

Results

Differences in dataset construction altered analysis populations, measurement timing, and endpoint composition, leading to distinct interpretations of treatment benefit despite identical statistical models. In some cases, reconstructed visit timing shifted subjects between response categories in others, alternative baseline definitions changed the magnitude of treatment effects. These findings indicate that the realised estimand is often operationalised through data architecture rather than modelling alone.

Discussion

Our results highlight that data management, statistical programming, and statistical design jointly define the inferential framework of a trial. Consistent with ICH E9(R1), the estimand is not fully specified until the pathway from raw data to analysis dataset is explicit. Greater transparency around data lineage and derivation logic may improve reproducibility, facilitate regulatory interpretation, and strengthen confidence in clinical trial conclusions.

Generating High-Fidelity Digital Twins: A Rigorous Statistical Framework for Privacy-Preserving Synthetic Biomedical Data

11:45-12:00

Presenter: EMİNE Kanik

Authors: EMİNE Kanik; Burak Akicier

Affiliations: AB Health Tech MedicReS, Houston

Introduction

In the era of biomedical AI, the demand for large-scale, granular datasets is at an all-time high. However, data privacy regulations and ethical constraints often limit the accessibility of real-world patient data. This study presents the E-PICOS Synthetic Data Module, a platform designed to generate 'Digital Twins' - statistically identical but privacy-protected synthetic datasets - that enable robust AI training and valid statistical inference without compromising patient confidentiality. Objective: To evaluate the statistical fidelity and utility of Digital Twins generated via the E-PICOS framework. We focus on the preservation of multivariate covariance structures, effect sizes, and distributional properties critical for biomedical research and power analysis. analysis.

Methods

We conducted a validation study using a mixed-variable dataset (N=100) containing 10 interdependent biomedical variables (e.g., Age, BMI, Systolic BP, Treatment, Outcome). The E-PICOS module utilizes a hybrid generation approach:

1.Covariance Preservation: For continuous variables, Cholesky decomposition is employed to replicate the exact correlation matrix and multivariate normal structures.

2.Distributional Fitting: Categorical and binary variables are modeled using automated distribution detection (e.g., Multinomial, Binomial) to maintain proportional accuracy.

3.Relationship Modeling: The system maintains complex inter-variable dependencies, including non-linear relationships and interaction effects.A Digital Twin dataset of N=1,000 was generated and compared against the original sample using a comprehensive validation suite.

Results

The generated Digital Twins demonstrated exceptional statistical fidelity:

•Correlation Preservation: The original correlation structure was maintained with 100% accuracy across all continuous variable pairs.

•Distributional Accuracy: Univariate statistics (Means, SDs, and Proportions) showed a 94% quality score, with minimal Standardized Mean Differences (SMD).

•Effect Size Stability: Critical effect sizes, including Cohen's d and Odds Ratios (ORs), were preserved within a 90% confidence interval of the original estimates.

•Overall Quality: The platform assigned an overall Quality Score of 90% (Excellent), indicating high utility for downstream AI model development and statistical testing.

Discussion

The E-PICOS Synthetic Data Module provides a statistically rigorous framework for generating high-fidelity Digital Twins. By preserving the underlying data generating process (DGP) and covariance structures, this technology facilitates a paradigm shift in biomedical research - enabling open science, multi-site collaboration, and AI innovation while ensuring absolute privacy compliance.

Evaluating Synthetic Patient Data for Robustness Assessment in Real-World Treatment Effect Estimation: A Simulation Study

12:00-12:15

Presenter: Paolo Messina

Authors: Paolo Messina; Sofia Azevedo; Mario Torchia; Luca Emili

Affiliations: InSilicoTrials Technologies Spa, Trieste

Introduction

Secondary data are increasingly used to generate real-world evidence, with improved analytical methods like target trial emulation guiding causal inference. These data often cover wider populations than the target trial's inclusion/exclusion (I/E) criteria, and rather than ignoring this, the broader data can enhance treatment effect estimates. Similar to Bayesian approaches, extra data can be incorporated without explicit prior settings, handling high-dimensional or non-linear covariate relationships.

Objectives

We evaluate a Synthetic Patient Data (SPD) framework that synthesizes from the full secondary data source before applying I/E criteria, and compare it with bootstrap resampling of the eligible cohort alone, using a simulation study with known ground truth.

Methods

Secondary data sources of n = 1,000 patients were simulated (binary treatment continuous outcome true ATE = +5.0 four confounders), of whom 649 (64.9%) met the I/E criteria of the target analysis. Three strategies were compared using inverse probability treatment weighting (IPTW): (i) single IPTW estimate on the eligible cohort (ii) bootstrap resampling of the eligible cohort across K = 500 replicates and (iii) SPD generated from the full secondary data source via fully parametric synthesis, with I/E applied post-synthesis (K=500). All strategies used the same IPTW estimator SPD datasets passed pre-specified fidelity and privacy thresholds before analysis.

Results

SPD outperformed bootstrap across all metrics. The mean SPD IPTW-ATE was 4.61 (bias = -0.39 relative bias = -7.9% RMSE = 0.94) versus 3.98 for bootstrap (bias = -1.02 relative bias = -20.5% RMSE = 1.37). Nominal 95% CI coverage was 95.4% for SPD versus 84.4% for bootstrap, demonstrating that bootstrap CIs are anti-conservative when I/E filtering truncates the covariate distribution. Results were consistent in both null, negative ground-truth, and confounding scenarios.

Table 1. Head-to-head performance: SPD vs bootstrap (true ATE = +5.0) Metric True ATE Single RWD Bootstrap (K=500) SPD (K=500) Winner Mean ATE +5.00 3.93 3.98 4.61 SPD Bias 0.00 -1.07 -1.02 -0.40 SPD Relative bias 0.0% -21.3% -20.5% -7.9% SPD RMSE - - 1.366 0.944 SPD Coverage of true ATE - n/a 84.4% 95.4% SPD 95% CI - 2.01-5.86 2.27-5.85 2.97-6.43 SPD CI = confidence interval IPTW = inverse probability of treatment weighting SPD = synthetic patient data.

Discussion

By synthesizing from the full secondary data source, SPD implicitly leverages the distributional breadth of the broader population, recovering information that bootstrap resampling, confined to the eligible cohort, necessarily discards. This yields substantially lower bias (relative bias -7.9% vs. -20.5%) and correctly calibrated CI coverage (95.4% vs. 84.4%). SPD offers a scalable, principled mechanism for borrowing external distributional strength, particularly valuable when secondary data sources are richer than their eligible subsets suggest.

Configurational Pathways to Pathological Complete Response (pCR) in Triple-Negative Breast Cancer (TNBC): A Robustness Assessment of Fuzzy-set Qualitative Comparative Analysis (fsQCA)

12:15-12:30

Presenter: Stefania Lando

Authors: Stefania Lando [1, 2, 3, 4]; Gloria Brigiari [1, 3, 4]; Ester Rosa [1, 3, 4]; Davide Massa [2, 5]; Valentina Guarneri [2, 5]; Maria Vittoria Dieci [2, 5]; Dario Gregori [1, 3]

Affiliations: [1] Unit of Biostatistics, Epidemiology and Public Health, Department of Cardiac, Thoracic, Vascular Sciences and Public Health, University of Padova, Padova; [2] Department of Surgery, Oncology and Gastroenterology (DiSCOG), University of Padova, Padova, Italy, Padova; [3] BIOSTAT-X Biostatistics & AI for Biomedical Discovery, Institute of Pediatric Research (IRP) 'Città della Speranza', Padova, Italy, Padova; [4] PhD Program in Translational Specialistic Medicine "G.B. Morgagni", Curriculum "Biostatistics and Clinical Epidemiology", University of Padova, Padova; [5] Oncology 2, Veneto Institute of Oncology IOV IRCCS, Padova, Italy, Padova

Introduction

Patients with early-stage TNBC are increasingly treated with pre-operative systemic therapy. Achieving a pCR, a complete tumor regression on the tumor bed and regional lymph-nodes, constitutes a key prognostic factor. While multiple clinical and biological variables contribute to treatment response, conventional regression models focus on independent effects and may overlook complex interactions. fsQCA provides an alternative framework to identify combinations of conditions associated with outcomes.

Methods

We applied fsQCA to identify combinations of conditions associated with pCR in TNBC. Clinical and biological variables, including tumor proliferation (Ki67), tumor-infiltrating lymphocytes (TILs), and treatment with immune checkpoint inhibitors (ICI) in addition to standard neoadjuvant chemotherapy, were calibrated into fuzzy-set memberships. We constructed truth tables and performed sufficiency analysis to derive configurational pathways leading to pCR, using standard inclusion and frequency cut-offs.

To assess methodological robustness, we conducted sensitivity analyses across alternative calibration schemes, evaluating how threshold specification affects the number, structure, and interpretability of solutions. We quantified calibration stability by examining the recurrence of solution terms across specifications. In addition, we performed bootstrap resampling to assess pathway stability under sampling variability. For comparison, we fitted a logistic regression model using the same predictors and evaluated its discriminative performance via area under the ROC curve (AUC).

Results

Sufficiency analysis identified a dominant configurational pathway to pCR defined by the joint presence of high Ki67, high TILs, and ICI treatment (consistency = 0.823, coverage = 0.140). This pathway

was reproduced across all alternative calibration schemes and recurred in 71.6% of bootstrap samples, indicating a robust core configuration. Logistic regression confirmed that Ki67, TILs, and ICI were independently associated with pCR (AUC = 0.713), but did not capture their joint configurational structure.

Discussion

The achievement of pCR in TNBC is associated with the joint presence of tumor proliferation, immune activation, and immunotherapy exposure rather than with single predictors alone. fsQCA complements regression-based approaches by identifying interpretable configurational pathways and by making robustness assessment central to inference.

Health Technology Assessment

Tue, 29.09.26, 11:00–12:30, Room: N4

Chair: Gaby Sroczynski

Meta-subgroup analyses at IQWiG: A comparison of the current statistical methodology with candidate methods for the future

11:00-11:15

Presenter: Frank Weber

Authors: Frank Weber; Ulrich Grouven; Jona Lilienthal; Guido Skipka; Ralf Bender

Affiliations: Institute for Quality and Efficiency in Health Care (IQWiG), Cologne

Introduction

At the Institute for Quality and Efficiency in Health Care (IQWiG), the current statistical methodology for analyzing how a treatment effect varies across subgroups of patients by synthesizing evidence from multiple trials (i.e., for 'meta-subgroup analyses') consists of conducting separate meta-analyses within the subgroups and then contrasting their results via the Q-test. Recently, Godolphin et al. (2022) proposed to take a different approach where the subgroups are contrasted first within each trial and then these within-trial contrasts are pooled, followed by an adjusted estimation of the subgroup-specific mean treatment effects. This procedure results in a 'trial-first' approach, whereas the current methodology at IQWiG can be considered a 'subgroup-first' approach. A midway approach is a multivariate meta-analysis where the off-diagonal elements of the between-trial covariance matrix are not fixed to zero.

Methods

To identify the most appropriate method for meta-subgroup analyses at IQWiG, we conduct a simulation study, starting with normal-approximation ('z') variants, but also aiming at variants similar to the Hartung-Knapp/Sidik-Jonkman (HKSJ) approach for standard meta-analyses in future investigations. The simulation scenarios are chosen such that they reflect meta-subgroup analyses at IQWiG as well as possible, with a limitation to two subgroups.

Results

Preliminary simulation results for the z-variants of the three approaches indicate that a trade-off needs to be made: The trial-first approach performs best in terms of coverage and type I error, but performs worst in terms of power, interval length, and mean squared error.

Discussion

Due to the results for the z-variants, HKSJ-type variants should be investigated as well. These (especially when combined with an ad-hoc variance modification) might mitigate under-coverage and inflation of type I error, at the cost of increased interval length and reduced power. Since the downsides of HKSJ-type variants are probably more pronounced in case of few trials, a possible further route could be Bayesian variants (of the subgroup-first and the midway approach) with strong priors for hierarchical variance parameters.

References

Godolphin PJ, White IR, Tierney JF, Fisher DJ (2022) Estimating interactions and subgroup-specific treatment effects in meta-analysis without aggregation bias: A within-trial framework. Research Synthesis Methods 14(1):68-78. https://doi.org/10.1002/jrsm.1590

Surrogate validation for health technology assessment: A simulation study comparing bivariate meta-analysis and the surrogate threshold effect

11:15-11:30

Presenter: Gregor Buch

Authors: Gregor Buch; Christian Stock

Affiliations: Boehringer Ingelheim Pharma GmbH & Co. KG, Ingelheim, Germany, Ingelheim am Rhein

Introduction

Using surrogate endpoints in clinical studies can accelerate patient access to new therapies. However, their statistical validation remains challenging. Within European joint clinical assessments (EU JCA), bivariate meta-analysis (BMA) and the surrogate threshold effect (STE) are considered for achieving the highest level of validation. A systematic comparison seems warranted to assess their performance.

Methods

A simulation study was conducted to evaluate how frequently surrogate validation is achieved by different validation methods. Study-level treatment effects on the linear predictor scale were generated under varying conditions, including the number of studies, the between-study correlation, between-study heterogeneity, and effect sizes. The BMA approach was implemented in a Bayesian framework validation was considered achieved when the posterior probability that the between-study correlation was at least 0.85 reached either 50% (BMA-50%) or 80% (BMA-80%). For the STE approach, validation was considered achieved when the confidence interval of a study's effect estimate exceeded the STE derived from meta-regression using either 95% or 80% prediction intervals (STE-95% and STE-80%).

Results

Validation occurred most frequently with BMA-50% and with STE-80%. In settings with only five studies, low heterogeneity, or small effect sizes, STE validated more often than BMA, regardless of the true underlying correlation. Notably, STE produced >5% validations even when the true correlation was zero (false-positive rate). BMA-50% validated most frequently with 15 studies and high true correlations between surrogate and true endpoint, as well as in scenarios with high heterogeneity combined with high true correlations. In the latter, BMA-80% occasionally validated more frequently than STE. BMA-50% validated most often in low effect size settings, independent of the true correlation. With 10 studies and moderate heterogeneity and effect size, STE-95% and BMA-50% showed comparable validation rates across correlation levels.

Discussion

Surrogate validation methods perform differently across evidence settings, highlighting the importance of aligning method choice with data scenarios. In some settings, false-positive validations were frequent - an issue critical for method selection but not addressed in guidelines. BMA controlled false-positive rates well, yet clear validation criteria are lacking. Currently, national assessment authorities involved in EU JCA provide explicit criteria only for STE, indicating an opportunity for refined EU JCA recommendations.

Modelling heterogeneous treatment effects in cost-effectiveness analysis using baseline risk scores

11:30-11:45

Presenter/Author: Konstantina Chalkou, Tasnim Hamza

Affiliations: Institute of Social and Preventive Medicine, University of Bern, Bern; F. Hoffmann-La Roche, Ltd, Basel

Introduction

Cost-effectiveness analyses often overlook heterogeneous treatment effects, potentially masking differences in treatment benefits across subgroups. Network meta-analysis allows for the synthesis of multiple treatments while accounting for effect modification by patient characteristics, when individual patient data are available. We present a methodological framework that integrates NMA-estimated heterogeneous treatment effects in cost-effectiveness analysis using a risk-modelling approach.

Methods

We used a previously developed two-steps prediction model based on network meta-regression. In the first step, a baseline risk score of the outcome is developed using prognostic factors in the second step, the baseline risk score is used in network meta-regression as the only effect modifier. Then we compute incremental cost-effectiveness ratios (ICER) and net monetary benefits (NMBs)

as functions of the baseline risk. We illustrate the approach using individual participant data from observational and randomised studies in relapsing-remitting multiple sclerosis, comparing dimethyl fumarate, glatiramer acetate, and placebo.

Results

ICER and NMB were changing with different baseline risk, so choosing the optimal drug is conditional to subpopulations. Dimethyl fumarate showed greater effectiveness in reducing disease progression hazard than glatiramer acetate overall, but its relative benefit declined with increasing baseline risk. The ICER of dimethyl fumarate versus glatiramer acetate increased from approximately 50,000 CHF per quality-adjusted life year (QALY) at 10% baseline risk to 210,000 CHF at 80%. Dimethyl fumarate has a higher NMB up to a baseline risk of 55%, after which glatiramer acetate becomes the preferred option. Population-level benefits were observed when comparing individualized treatment strategies to standard, average-based approaches.

Discussion

Our approach shows the importance of accounting for heterogeneous treatment effects in cost-effectiveness evaluation. We present a framework that combines risk modelling (an approach widely used in prognostic research) in combination with network meta-regression and cost-effectiveness analysis to provide individualised cost-effectiveness information.

Estimating health-state utility for economic models in event-driven clinical studies with prevalence of death - A case study based on a relevant causal estimand and a robust and efficient estimator

11:45-12:00

Presenter: Robert Bauer

Authors: Robert Bauer [1]; Margarida Augusto [1]; Antonio Remiro Azócar [1]; Julie Funch Furberg [1]; Martin Bøg [1]; Klaus Kähler Holst [1]; Christian Bressen Pipper [1, 2]

Affiliations: [1] Novo Nordisk A/S, Søborg; [2] University of Southern Denmark, Odense

Introduction

Health-state utility data are of critical importance in cost effectiveness models to determine quality-adjusted life years. Whereas general guidance around health-state utility estimation is available (ISPOR 2016), detailed guidance is lacking around utility estimation based on event-driven clinical studies, where terminal events such as death are prevalent.

Methods

We discuss challenges related to the application of existing approaches including a) hypothetical estimands to handle the intercurrent event of death and b) composite strategies where unfavourable

utility values are assigned. Instead, we propose targeting a more natural causal estimand evaluating the treatment effect conditional on no terminal event. The proposed methodology (Holst et al. 2025) is based on a partly conditional approach and enables a robust, efficient, and assumption lean assessment of the effect of treatment interventions simultaneously on both the health-state utility score and the onset of a terminal event at a landmark time.

Results

In our case study we apply the methodology to FLOW, a large-scale event-driven randomized controlled study in a population with type 2 diabetes and chronic kidney disease at high risk of kidney disease progression (Perkovic et al. 2024). EQ-5D data were collected in FLOW and corresponding utilities mapped based on UK tariffs. Health-state utility estimation results at landmark year 2 after randomization are provided based on a) naïve estimates and b) covariate adjustment with the latter showing considerable efficiency gains in terms of precision.

Discussion

In conclusion, this case study shows the applicability of the Holst methodology to the health-state utility estimation context, providing an interpretable, robust, and efficient approach to estimate health-state utilities in event-driven clinical studies.

References

Holst et al. 2025:

https://arxiv.org/pdf/2502.03942

ISPOR 2016:

https://pubmed.ncbi.nlm.nih.gov/27712695/

Perkovic et al. 2024:

https://www.nejm.org/doi/full/10.1056/NEJMoa2403347

PAC-Bayesian Reward-Certified Outcome Weighted Learning

12:00-12:15

Presenter: Yuya Ishikawa

Authors: Yuya Ishikawa [1, 2]; Shu Tamano [3, 4]

Affiliations: [1] National Cancer Center Japan, Tokyo; [2] Institute of Science Tokyo, Tokyo; [3] The University of Tokyo, Tokyo; [4] Japan Institute for Health Security, Tokyo

Introduction

Outcome weighted learning (OWL) estimates individual treatment rules by framing value maximization as a weighted classification problem. In many applications, however, the observed reward is

often just a proxy for the target utility of interest. It may systematically overestimate this utility, leading to optimistic biases in policy learning. We study OWL when the utility is not fully observed and establish finite-sample guarantees for the true target value.

Methods

We propose PAC-Bayesian Reward-Certified Outcome Weighted Learning (PROWL) for policy learning under reward uncertainty. Under a one-sided bound on how much the observed reward can exceed the target reward, we show that the certified value equals a policy-independent constant minus a cost-sensitive classification risk. This reduction allows nuisance and policy parameters to be learned jointly without sample splitting and yields a nonasymptotic PAC-Bayes lower bound on the target value for randomized treatment rules. We further show that the posterior distribution that maximizes this bound coincides with a general Bayes posterior and resulting Gibbs policy serves as our proposed PROWL estimator. For practical settings, we also derive an automated, data-driven rule to select the learning rate based on these bounds and introduce a computationally efficient surrogate posterior using a hinge loss relaxation. The associated surrogate loss is shown to be Fisher consistent for the certified classification target.

Results

Our method provides three main theoretical results. First, the certified value provides a valid lower bound on the unobserved target value, while its exact reduction to a classification risk enables split-free joint learning of nuisance and policy components. Second, the PAC-Bayes provides a finite sample lower bound, and the optimal posterior is characterized by a general Bayesian update. Third, the same lower bound principle naturally yields automatic, data-driven learning-rate calibration. Across simulation settings with reward uncertainty, PROWL achieves lower regret in target value than conventional methods.

Discussion

PROWL extends outcome weighted learning from the direct optimization of proxy rewards to the certified, theoretically guaranteed optimization of an unobserved true utility. By integrating exact value certification, PAC-Bayes lower bounds, and a Fisher-consistent surrogate, this framework provides a theoretical and practical approach to learning individual treatment rules under reward uncertainty.

Bridging the Gap: Real-World Safety Assessment to Enable Inclusive Clinical Trials

12:15-12:30

Presenter: Rana Jreich

Authors: Rana Jreich [1]; Sacha Levy [2]; Kahina Lounaci [2]; Helgi Van de Velde [3]; Zhaoling Meng [4]

Affiliations: [1] Sanofi, Gentilly; [2] Quinten Health, Paris; [3] Sanofi, Ghent; [4] Sanofi, Boston

Introduction

Clinical trial populations often fail to reflect real-world patients due to restrictive - and sometimes unnecessarily restrictive - eligibility criteria. While broadening eligibility criteria can improve trial generalizability, doing so requires assurance that safety is not compromised. Real-world evidence, derived from heterogeneous patient populations treated in routine practice, provides an opportunity to evaluate the safety implications of expanding eligibility. Although several studies have examined the impact of eligibility criteria on efficacy, limited work has focused specifically on safety, highlighting the need for structured approaches to assess risks in previously excluded subgroups.

Methods

We developed a quantitative real-world data (RWD) framework to evaluate safety risks associated with patient subgroups commonly excluded from clinical trials. Rather than analyzing criteria individually - which often limits statistical power and interpretability - we focused on clinically meaningful subpopulations defined through expert. For each subgroup, safety outcomes were assessed by treatment class. Confounding was addressed using inverse probability of treatment weighting to balance baseline characteristics between patients with and without baseline condition. Weighted Kaplan-Meier analyses were then applied to compare safety event risks. This framework was implemented in patients with relapsed/refractory multiple myeloma receiving standard-of-care (SoC) therapies or anti-CD38 treatments.

Results

RRMM patients with renal impairment showed significantly higher rates of severe renal events, without corresponding increases in severe hematologic events or infections. Patients with pre-existing cardiac dysfunction exhibited increased risks of infections and cardiovascular events and also experienced higher rates of severe renal events under SoC therapy. Furthermore, patients who had received more than two prior lines of therapy demonstrated elevated risks of severe hematologic events, infections, and cardiovascular complications. These risks were consistently higher in the SoC cohort compared with patients treated with anti-CD38 therapies, suggesting differential vulnerability depending on treatment class.

Discussion

These findings demonstrate the value of RWD in characterizing safety across diverse patient groups. By quantifying risks in subpopulations traditionally excluded from trials, this framework supports evidence-based expansion of eligibility criteria. This enables more inclusive trial designs while maintaining essential safety standards and improving the relevance of clinical research to real-world patients.

Strategies for Prediction Models

Tue, 29.09.26, 11:00–12:30, Room: Runder Saal

Chair: Richard Riley

Recent criteria and simple rules for the required sample size to develop clinical prediction models: theoretical and empirical comparisons

11:00-11:15

Presenter: Ewout Steyerberg

Authors: Ewout Steyerberg [1]; Toby Hackmann [2]; Ben Van Calster [3]; Maarten van Smeden [1]; Laure Wynants [4]

Affiliations: [1] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht; [2] Leiden University Medical Center, Leiden; [3] Department of Development and Regeneration, KU Leuven, Leuven; [4] Maastricht University, Maastricht

Introduction

Adequate sample size is essential to the development of prediction models with binary outcomes. We aim to relate recent approaches to traditional rules of thumb ('simple rules') and assess agreement and differences in judging adequacy of such sample size.

Methods

A well-known simple rule considers the events per variable (EPV), or specifically, events per predictor parameter p (EPP, e.g. 'EPP>10', or 'EPP>20') to limit overfitting to small data sets. Another rule requires a minimum number of events (E) for reliable estimation of the overall event rate (e.g. 'E>100'). Recent criteria consider: 1) limited overfitting in predictor effect estimates ('global shrinkage s $\geq$ 0.9') 2) small optimism in Nagelkerke's R2 ('delta(R2) $\leq$ 0.05') and 3) precise estimation of the overall event rate (margin of error <0.05). We use statistical theory to compare simple rules to these three recent criteria.

Results

Simple rules and recent criteria 1 and 2 compare well for relatively rare events, where the number of events defines the effective sample size. Global shrinkage relates model fit (chi2, equivalent to Area under the ROC curve c for a given event rate) to model complexity (parameters p): s=(chi2 - p) / chi2. At a 10% event rate, criterion 1 appears to correspond to the classic EPP>10 rule for an Area under the ROC curve (c) of 0.767, and EPP>20 for c= 0.694. Criterion 2 implied EPP>4 (largely irrespective of c). Criterion 3 considers the absolute error, while E=100 is approximately a 10% relative error, with agreement between criterion 3 and E=100 at a 30% event rate. Where recent criteria may be combined in a max function, we also consider a rule for the number of events as 100+10*p. We

found agreement for EPP>10 and especially for this combination rule with max(recent criteria) in a comparison of required sample size for 299 published COVID-19 prediction models.

Discussion

Simple rules have direct relations with recent criteria for sample size calculations to limit overfitting and to guarantee reliability of predictions. Recent criteria are essential to inform prospective prediction modeling efforts, while simple rules may still be sufficient to help appraise the quality of already developed prediction models.

Estimating Stability of Person-Specific Predictions using Bootstrapping

11:30-11:45

Presenter: Nicole Erler

Authors: Martijn van Dam [1]; Alex Carriero [1]; Nicole Erler [1]; Anne de Hond [1]; Karel Moons [2]; Richard Riley [3]; Gary Collins [3]; Maarten van Smeden [2]

Affiliations: [1] Department of Data Science & Biostatistics, Julius Center, University Medical Center Utrecht, Utrecht; [2] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht; [3] University of Birmingham, Birmingham

Introduction

Clinical prediction models generate person-level risk predictions that are subject to uncertainty. Riley and Collins (2023, Biom. J., 65, 220302) proposed estimating the uncertainty arising from model training by studying the stability of person-level predictions obtained from repeatedly retraining the model on bootstrap samples. So far, the performance of bootstrapping for evaluating stability, rather than standard measures of predictive performance, has not been thoroughly evaluated.

Methods

To evaluate the performance of bootstrapping to assess the stability of person-level predictions, we conducted a simulation study in which we trained predictions models using Elastic Net logistic regression, unpenalized logistic regression, and Extreme Gradient Boosting trees (XGBoost) with varying sample sizes. We compared non-parametric bootstrap against 10-fold cross-validation (CV) and out-of-bag bootstrapping as practical alternatives. As a reference, we compared these procedures to independent random samples from the target population, as generated from the data-generating mechanism on which the prediction models are trained. The performance of the resampling methods is compared using summaries of the resulting distribution of predictions to corresponding summaries from independent sampling. We studied and compared the variances of the risk predictions as well as the Mean Average Prediction Error (MAPE) with respect to the initial predictions and the MAPE of the 'true' risk scores.

Results

Preliminary results indicate that the original bootstrap performs suboptimally for XGBoost and to a lesser extent for Elastic Net logistic regression. Performance of stability evaluation can be substantially improved by using the out-of-bag predictions from bootstrapping for XGBoost. More extensive simulation results will be presented at the conference.

Discussion

The results of our simulation indicate that the standard bootstrapping procedure for studying stability of prediction models may vary depending on the methods used to develop the prediction model. These results can be explained by the mechanisms by which XGBoost and Elastic Net arrive at their predictions.

Conformal prediction under miscalibration: what coverage and interval width do not capture

11:45-12:00

Presenter: Giulia Zamagni

Authors: Giulia Zamagni [1]; Giulia Barbati [2]

Affiliations: [1] Institute for Maternal and Child Health - IRCCS 'Burlo Garofolo', Trieste; [2] University of Trieste, Trieste

Introduction

Conformal prediction is increasingly used for uncertainty quantification in machine learning, as it provides distribution-free guarantees on marginal coverage of the prediction set. However, in clinical settings, predicted probabilities are often used directly for decision-making, requiring not only valid coverage but also well-calibrated risk estimates. It remains unclear whether conformal outputs, including Venn-Abers (VA) predictors, meaningfully reflect model calibration in practice.

Methods

We conducted a simulation study considering true risks as known by construction. Miscalibration was induced by inflating the linear predictor on the logit scale (model A), while improved calibration was obtained via global shrinkage (model B). Split conformal classification was applied to construct prediction sets using non-conformity scores derived from predicted probabilities of the observed class (nominal coverage 0.95 calibration sample size n=100). Efficiency was quantified by the singleton rate. We further evaluated VA predictors, which apply isotonic mapping to predicted probabilities to produce probability intervals. To isolate the role of the model calibration, we reduced the calibration set for model B (model C n=50), yielding a weakly determined isotonic map.

Results

Non-conformity score distributions differed markedly across models, being more asymmetric under miscalibration (model A). While both models A and B achieved nominal coverage (0.95 and 0.96), model A yielded a lower conformal threshold and higher efficiency (singleton rate 0.44 vs 0.40), reflecting sharper but overconfident predictions. For VA predictors, interval-width distributions were nearly identical for models A and B when the calibration sample was large (n=100). In contrast, reducing the calibration sample (model C) led to highly heterogeneous interval widths, indicating sensitivity to the stability of the isotonic calibration map.

Discussion

Models with markedly different calibration can yield similar conformal coverage and VA interval widths, indicating that these quantities are insensitive to substantial differences in model calibration. Moreover, VA intervals depend strongly on the stability of the isotonic calibration map, which may be weakly determined in finite samples. These findings highlight that, when predicted probabilities are used for clinical decision-making, reliable use of such predictions requires well-calibrated underlying models. Therefore, calibration remains a separate and essential requirement.

Incorporating prognostic scores for non-collapsible time-to-event endpoints

12:00-12:15

Presenter: Harry Parr

Authors: Harry Parr; Douglas Thompson; Tasos Papanikos; Aris Perperoglou

Affiliations: GSK, London

Introduction

Clinical trial analyses can be made more efficient when we suitably leverage baseline prognostic information via covariate adjustment. The increased availability of observational and historical clinical trial data has promoted new methodological opportunities in how the value of existing data may be exchanged within a new trial setting. One such area is in the use of prognostic scores, where prognostic information is combined into an overall derived score. The benefits and properties of these scores are well-founded, namely improving precision leading to greater statistical power and hence ultimately greater probability of regulatory success. The primary aim of this work is to explore a route toward a unified framework: simultaneously "borrowing" from historical data via prognostic scores which can then be leveraged under modern causal estimation methods (e.g. g-computation).

Methods

Inference on non-collapsible summary measures is challenging in covariate adjustment, e.g. odds or hazard ratio endpoints. These move further away from the null in covariate adjustment the increases seen in power are faux in that no further precision gains are made in the effect estimate. The resulting

marginal and conditional treatment effect estimates are not equivalent. We designed a time-to-event simulation to study how the adoption of causal methods offers a solution to this challenge specifically using g-computation. We consider this relative to alternative summary measures and estimands, such as the restricted mean survival time (which are collapsable).

Results

We investigate the properties of prognostic scores under time-to-event outcomes via simulations, in simple cases of no censoring, and non-informative censoring mechanisms. We generate prognostic scores of varying strengths (from nil to strong) to assess impact on power and HR estimates to measure (any) precision gain in the treatment effect. Work is ongoing and will be presented in full.

Discussion

Extending the application of prognostic scores to inform of contemporaneous trial design is well known, though there have been relatively few applications in nonlinear/time-to-event endpoints. Here we provide a framework to incorporate conditional prognostic scores into time-to-event analyses, providing recommendations for estimation and quantifying the expected precision gains in the treatment effect.

Feedback Loops in Clinical Prediction Models: A scoping review

12:15-12:30

Presenter: Ziyi Ding

Authors: Ziyi Ding; Haya Elayan; Brian McMillan; Jamie Sergeant; David Jenkins

Affiliations: The University of Manchester, Manchester, UK, Manchester

Introduction

Following deployment, clinical prediction models (CPMs) undergo monitoring, evaluation, and updating based on deployment data. However, the performative effects of deployed CPMs are often ignored, giving rise to the feedback loop problem. Here, deployed CPMs are maintained using data altered by their own predictions, undermining the validity of post-deployment evaluation and updating, with implications for patient safety and outcomes. The existing literature remains fragmented across terminology, disciplines, and stages of the model lifecycle. We therefore conducted a scoping review to map the current state of this topic, identify gaps and establish a foundation for future research.

Methods

We conducted keyword searches in Ovid (MEDLINE and Embase) and Google Scholar, complemented by forward citation tracking. We included studies discussing feedback loops in CPMs and excluded studies on generic AI or decision support systems without an explicit clinical prediction

focus. Data were extracted on study characteristics, feedback loop pathways, consequences, terminology, and mitigation strategies and their applicability and limitations, and synthesised narratively to characterise the structure of the topic.

Results

A total of 51 studies were included in the final analysis. We identified (1) two feedback loop pathways (outcome-mediated and observation-mediated) and five associated scenarios (2) possible consequences of the feedback loop, including two paradoxical patterns in model assessment and three types of biased updating outcomes (3) inconsistency and lack of standardisation in terminology and (4) a range of proposed mitigation strategies, with causal inference-based approaches the most common however, each proposed strategy was subject to methodological, ethical, or practical limitations, with limited implementation in practice.

Discussion

This scoping review synthesises multiple heterogeneous aspects of feedback loops in CPMs and suggests more consistent terminology. More importantly, it highlights the lack of consensus and guidance on mitigation strategies, which also remain under-recognised within current regulatory frameworks, underscoring the need for future research to generate empirical evidence to inform strategy decision. Finally, the feedback loop problem should be understood as a system-level issue requiring coordinated efforts, including improvements in data management and deployment infrastructure.

13:30–15:00

Data Quality, Selection, and Measurement

Tue, 29.09.26, 13:30–15:00, Room: K1

Chair: Apostolos Gkatzionis

Application of Data Quality Summary Reports (DQ-SR) at one MII MeDIC - A Technical Case Report

13:30-13:45

Presenter: Christian Draeger

Authors: Christian Draeger [1]; Dennis Hübner [2]; Sherry Freiesleben [3]; Alexander Kiel [4]; Julian Saß [5]; Margaux Gatrio [6]; Philip Overton [7]; Thomas Kleinert [7]; Matthias Löbe [4]

Affiliations: [1] Universität Leipzig: Institut für Medizinische Informatik, Statistik und Epidemiologie, Leipzig; [2] Thiem Research GmbH, Cottbus; [3] Rostock University Medical CenterUniversitätsmedizin Rostock, Rostock; [4] Universität Leipzig, Leipzig; [5] Berliner Institut für Gesundheitsforschung in der Charité - Universitätsmedizin Berlin, Berlin; [6] BIH Charité, Berlin; [7] TMF e.V., Berlin

Introduction

The German Medical Informatics Initiative (MII) aims to enable secondary use of clinical data by providing data from university hospitals in a common data model based, the MII Core Dataset (CDS), based on HL7 FHIR. CDS adoption varies regarding the detail across sites due to differences in EHR systems, local infrastructures, and state regulations. This variation complicates the assessment of data readiness for research and the comparison of DQ between institutions. While standardized Data Quality Summary Reports (DQ-SRs) have been proposed to address this challenge, their practical implementation within existing data integration pipelines has yet to be explored.

Current data quality assessment approaches within the MII primarily focus on local, detailed analyses of DQ issues. Similarly, approaches from the OMOP community provide comprehensive assessment tools, but require additional data transformation. Produced reports containing variable-level information, limiting cross-site sharing. Consequently, these solutions do not meet the requirements for aggregatable, interoperable reporting defined by DQ-SRs.

Methods

To address these limitations, we implemented a FHIR-native approach for generating DQ-SRs as FHIR MeasureReport resources. Using the Blaze FHIR Server and its support for the Clinical Quality Language (CQL), data quality measures are specified as executable CQL scripts and evaluated directly on the FHIR server. A command-line tool automates the creation of FHIR Measure and Library

resources, orchestrate evaluation, and produce both machine-readable and HTML outputs, which are integrated into a dashboard.

Results

Deployment at a MeDIC with over 7.5 million instances of the Condition resource demonstrated that all CQL-based measures executed in under ten minutes per resource, while the flattening approach could not process the full dataset within a day. The resulting reports provide meaningful overviews of CDS harmonization and allow additional data quality checks beyond mandatory elements.

Discussion

Our experience shows that DQ-SRs are effective for cross-site harmonization assessment. Interpreting structural differences in FHIR resources still requires additional context, which is currently not machine-readable in the CDS IGs. Future improvements in metadata accessibility and machine-readable guides are essential to fully realize harmonization across sites. Such approaches may also serve as a model for multi-site healthcare data networks seeking efficient, scalable, and shareable quality assessment.

The full paper of this abstract will be published in the GMDS series "Studies in Health Technology and Informatics" with doi: 10.3233/SHTI260982, URL:https://dx.doi.org/10.3233/SHTI260982

Detecting systematic measurement changes during data collections

13:45-14:00

Presenter: Carsten Oliver Schmidt

Authors: Carsten Oliver Schmidt [1]; Ronja Foraita [2]; Elisa Kasbohm [3]; Elena Salogni [3]; Annika Swenne [2]

Affiliations: [1] Universität Greifswald, Greifswald; [2] Leibniz Institute for Prevention Research and Epidemiology - BIPS, Bremen; [3] Universitätsmedizin Greifswald, Greifswald

Introduction

Observational studies and clinical trials often collect data over several months or years, where examiners may perform differently, devices may age or be replaced, and environmental conditions may vary. This can introduce systematic measurement changes that, if ignored, may seriously bias study results. Unlike biological associations such changes generally lack any predictable direction or structure. This talk provides an overview on typically observed systematic measurement change from a population-based cohort study (Study in Health of Pomerania-SHIP [1]) and from a simulation study assessing the appropriateness of different statistical methods to detect them.

Methods

We used data from the SHIP-START cohort (baseline n=4,318). SHIP results motivated the choice of 132 simulation scenarios that varied in type of systematic change (jump, linear, quadratic, recurrent), distribution (normal, log-normal), signal-to-noise ratio, and magnitude of systematic error. The simulated sample sizes ranged from n=30 to n=1000). The simulation study followed a structured approach proposed by Morris et al. [2], studying the ability of seven statistical methods: Autoregressive integrated moving average (ARIMA), fused lasso signal approximator (FLSA), generalized additive model (GAM), locally weighted scatterplot smoothing (LOWESS), moving average (MA), piecewise regression (PR), and pruned exact linear time (PELT). Methods were compared according to their ability to detect and characterize systematic change, including its magnitude and the number of change points.

Results

In SHIP, diverse patterns of heterogeneity and potential measurement error were observed, ranging from gradual drifts to abrupt changes, occurring once or repeatedly. Applying the selected statistical methods to SHIP data and simulated data yielded markedly divergent results, highlighting how strongly outcomes depend on method choice. Across most simulation scenarios, GAM and LOWESS showed the most robust performance, whereas ARIMA and PELT were less reliable.

Discussion

Our findings demonstrate the importance of identifying methods that can reliably detect temporal measurement changes during data collection. They provide practical guidance for selecting appropriate approaches during data monitoring and Initial Data Analysis, guided by known data properties such as sample size and outcome distribution.

[1] DOI: 10.1093/ije/dyac034

[2] DOI: 10.1002/sim.8086

Adaptation and Validation of the Stanford GVHR for German-Speaking Countries: A Biometric Protocol for Multidimensional Gender Assessment

14:00-14:15

Presenter: Carlotta Derad

Authors: Carlotta Derad [1]; Charlotte Nettekoven [2, 3]; Stephanie Heinemann [4]; Pauline Kaboth [5]; Roland Gerard Gera [6]; Justyna Wünstel [7]; Ute Chiriac [8]; Vivian Kölbel [9]; Felicitas LaRosée [10]; Anne Christine Selbmann [10]; Gabriele Ihorst [11]; Monika Sadlonova [4, 12, 13, 14]; Christine A.F. von Arnim [4]; Christoph Herrmann-Lingen [15]; Martin Hellmich [16]; Heidrun Golla [2]

Affiliations: [1] University Medical Center GöttingenDepartment of Medical Statistics, Göttingen; [2] Department of Palliative Medicine, University Medical Center Göttingen, Göttingen; [3] University of

Cologne, Faculty of Medicine and University Hospital, Center for Neurosurgery, Department of General Neurosurgery, Cologne; [4] Department of Geriatrics, University Medical Center Göttingen, Göttingen; [5] Institute and Policlinic of Occupational and Social Medicine, Faculty of Medicine, Technische Universität Dresden, Dresden; [6] Carl von Ossietzky Universität Oldenburg, KKS / Abteilung Epidemiologie und Biometrie, Oldenburg; [7] Institute for General Practice, Goethe-Universität Frankfurt am Main, Frankfurt am Main; [8] Hospital Pharmacy, Heidelberg University Hospital, Heidelberg; [9] Clinic for General, Visceral and Transplant Surgery, Heidelberg University Hospital, Heidelberg; [10] Zentrum für Medizinische Informatik / Institut für Medizinische Informatik und Biometrie (IMB) Medizinische Fakultät Carl Gustav Carus, Dresden; [11] Zentrum Klinische Studien, Universitätsklinikum Freiburg, Freiburg im Breisgau; [12] Department of Cardiovascular and Thoracic Surgery, University Medical Center Göttingen, Göttingen; [13] Department of Psychosomatic Medicine and Psychotherapy, University Medical Center Göttingen, Göttingen; [14] Department of Psychiatry, Massachusetts General Hospital, Boston; [15] Universitätsmedizin Göttingen - Klinik für Psychosomatische Medizin und Psychotherapie, Göttingen; [16] Department of Medical Statistics, University Medical Center Göttingen, Göttingen

Introduction

Within the Gender Data Gap clinical research in Germany lack of standardized, validated instruments quantifying non-biological gender aspects. The Stanford Gender-Related Variables for Health Research (GVHR) offers a promising multidimensional approach, but its original validation is primarily within the Anglo-American context. This study adapts and validates the GVHR for the German-speaking population (GER-GVHR, DRKS00038648) and benefits from an interdisciplinary team spanning multiple discipline within medicine and biometry.

Methods

The GER-GVHR translation followed a standardized procedure (Beaton et al., 2000). Participants are recruited through email, bulletin board, and snowball sampling via private networks. Anonymized data is collected via REDCap survey. We aim for $n \geq 500$ to ensure adequate power for complex analyses, including Structural Equation Models. The validation pipeline comprises:

1. Confirmatory Factor Analysis: Assessment of the psychometric properties and model fit (CFI, TLI, and RMSEA) of the seven dimensions in the GER-GVHR.

2. Latent Profile Analysis: Identifying gender phenotypes beyond binary sex categories.

3. Associative Validation: Examining the relationship between GER-GVHR dimensions and established measures of mental health (PHQ-9), stress (PSS-10), and health-related quality of life (CDC HQoL-4).

Results

The translation and cultural adaptation of the GVHR have been completed. Data collection starts in April 2026. We plan to analyze the data this summer and present preliminary results at the conference.

Discussion

Validation of the GER-GVHR will provide a standardized, multidimensional instrument for gender assessment in the German-speaking world. This will facilitate more nuanced, gender-sensitive clinical and epidemiological research. We will present the validation protocol and preliminary results from this currently ongoing multi-phase survey.

References:

Nielsen, M.W., Stefanick, M.L., Peragine, D. et al. Gender-related variables for health research. Biol Sex Differ 12, 23 (2021). https://doi.org/10.1186/s13293-021-00366-3

Beaton, D. E., Bombardier, C., Guillemin, F., & Ferraz, M. B. (2000). Guidelines for the

process of cross-cultural adaptation of self-report measures. Spine, 25(24), 3186-

3191.

Representativeness is not always desirable: how selected samples may suffer less from unmeasured confounding

14:15-14:30

Presenter: Kate Tilling

Authors: Kate Tilling; Apostolos Gkatzionis; Rachael Hughes; Chin Yang Shapland

Affiliations: University of Bristol, Bristol

Introduction

Many medical research studies suffer from non-random participation or attrition. Such selection can lead to bias ("selection bias") when exposure (or a cause of exposure) and outcome (or a cause of outcome) cause selection. Awareness of this problem has increased over recent years, and has led some to assume that all analyses in all selected samples are biased.

Methods

We derive theoretical results for the cases where (i) there is an unmeasured confounder U, and the unmeasured confounder causes selection ii) additionally (but independently) a cause of exposure (V) and a cause of outcome (W) cause selection, and selection follows a log-additive model with interaction parameter (for V and W) $\delta 1$ (iii) additionally the log-additive model includes interaction terms $\delta 2$, $\delta 3$ for UW and UV. We use simulations to explore the impacts of different parameter choices.

Results

(i) We show that the bias (due to unmeasured confounding) in the selected sample is always in the same direction as the bias in the full sample - and where the selection reduces the variation in the unmeasured confounder, the bias is lower in the selected sample than the full sample (ii) where all variables are positively related, the bias in the selected sample can be reduced or increased in the

selected sample depending on the relative contributions of confounding and selection - specific parameters needed to assess the likely size and direction of bias (iii) where all variables are positively related, bias in the selected sample is less than that in the full sample where all $\delta i<0$.

Discussion

In some selected samples (e.g. UK Biobank, where selection has resulted in a sample that is healthier and wealthier than the UK population), bias due to unmeasured confounding might actually be lower than in the full population, where no other analysis variables are related to selection. In more complex scenarios, the likely direction of each selection mechanism can be used to assess whether the bias is likely increased or reduced in the selected sample. More representative samples will not always lead to less biased coefficients, in the presence of unmeasured confounding.

Addressing nonrandom nonresponse in survey-based causal inference using register data

14:30-14:45

Presenter/Author: Frank Eriksson

Introduction

Due to nonrandom inclination to participate in a survey, nonresponders may not exhibit the same cause and effect relations as responders. We generalize causal effects from survey responders to the full cohort by leveraging register-based shadow variables, available for all invited individuals, that provide information on unobserved responses.

This work is motivated by a Danish study on the side effects of hormonal contraception, where almost one million females aged 15-45 were invited to participate in a survey and near one-fourth responded. Rich data is available on the responders, including self-reported psychiatric history, lifestyle, sexual activity, relationship status and contraceptive use. By linking invited individuals to national registers, information such as medical history, prescriptions, residency and civil records are available on all invitees. The joint availability of survey and register data renders inference on contraceptive-related adverse events among responders plausible. Extending the scope to the full cohort is challenging if the readiness to participate in the study, and contribute with self-reported information, is related to the individual's response, had she chosen to respond.

Methods

Using methods developed for inference with data missing not at random, we investigate how register-based shadow variables can help recover the full-cohort effect. Shadow variables are associated with the missing data, but conditionally independent of the response mechanism given the full information. For example, when studying self-perceived depression, which may influence participation, antidepressant prescriptions can act as a proxy for depression that is unrelated to the propensity to participate once self-perceived depression is accounted for.

Results and Discussion

We describe settings where the outcome, exposure and/or confounders are self-reported and only observed among responders. For each scenario, we examine assumptions under which the causal effect in the full cohort is identifiable using shadow variables and propose corresponding estimation strategies.

A Two-Stage Structural Equation Modeling Framework to Address Spatial Confounding in Binary Response Models

14:45-15:00

This is one of this year's ISCB CASP award-winning abstracts.

Presenter: Md. Hasibul Islam Jitu

Authors: Md. Hasibul Islam Jitu; Syed Shahadat Hossain

Affiliations: Institute of Statistical Research and Training (ISRT), University of Dhaka, Dhaka

Introduction

Confounding occurs when a variable affects both predictors and outcome. When that variable also follows a spatial structure that is not properly adjusted for, the resulting bias is known as spatial confounding. This creates a methodological challenge because unmeasured spatial dependence can influence associations between covariates and outcomes. In many public health and social science settings, an additional complication is that key determinants, such as socioeconomic status or household environment, are latent rather than directly observed. Structural Equation Modeling (SEM)-based logistic regression models can address latent constructs but do not account for spatial dependence. Logistic regression also ignores spatial dependence and does not incorporate latent constructs, while spatial models such as the Conditional Autoregressive Generalized Linear Mixed Model (CAR-GLMM) capture spatial dependence but do not include latent constructs. As a result, no framework exists for binary outcomes that jointly models latent constructs and accounts for spatial confounding.

Methods

This study proposes a two-stage Spatial SEM (SSEM) framework that integrates confirmatory factor analysis for latent constructs with a GLMM containing a Conditional Autoregressive random effect to represent spatial dependence, enabling joint modeling of latent constructs while accounting for spatial confounding in binary outcomes.

Results

Simulation experiments across varying spatial dependence, sample sizes, and latent-structure strengths show SSEM consistently outperforms logistic regression and SEM and achieves predictive performance comparable to, and often slightly better than, the CAR-GLMM, particularly under stronger spatial dependence and larger samples. Supplementary simulations show that SSEM achieves lower bias and RMSE than the SEM, indicating improved accuracy. In the real-data application using the

Sample Vital Registration System, SSEM identifies key latent predictors of childhood deaths due to pneumonia and other respiratory diseases and achieves better predictive performance than the CAR-GLMM and non-spatial models. A de-spatialized predictor version is used to assess spatial confounding transmitted through indirect pathways. The stability of associations between the original and de-spatialized models suggests these indirect effects are weak, supporting robustness of the findings.

Discussion

This study introduces a framework that jointly models latent constructs and spatial effects, thereby accounting for spatial confounding, and achieves predictive performance and estimation accuracy comparable to or better than conventional approaches.

Statistical Learning under Uncertainty

Tue, 29.09.26, 13:30–15:00, Room: K2

Chair: Christoffer Lingjærde

Machine-learning Methods for Matched Case-Control Studies

13:30-13:45

Presenter: Gunther Schauberger

Authors: Gunther Schauberger; Stefanie J. Klug

Affiliations: TUM School of Medicine and Health, Technische Universität München, München

Introduction

Matched case-control studies are typically analysed using conditional logistic regression, while more flexible machine learning methods are rare. Regular machine learning methods cannot be applied directly to matched case-control data, as they do not account for the matching structure. Three new machine learning methods are presented that have been developed specifically for the analysis of matched case-control studies.

Methods

The three methods presented are a decision tree approach (CLogitTree) [1], a random forest approach (CLogitForest) [2], and a boosting approach (CLogitBoost). All three methods relax the strict linearity assumption of conditional logistic regression while still accounting for matching. For CLogitTree and CLogitForest, this is reached by embedding the respective trees into the framework of conditional logistic regression. For CLogitBoost, a new family was implemented in the R package mboost to account for the specific data structure of matched case-control data. In all methods, a distinct exposure variable can be modelled using a simple linear term for simple interpretation. In this case, the methods are used to more flexibly adjust for confounders and additional predictors.

Results

The methods are illustrated and compared using real-world data from a matched case-control study on the effect of regular screening attendance on cervical cancer, but also with simulated data. The results show that, depending on the data, all of them can reduce bias in estimating the exposure effect. CLogitTree and CLogitBoost provide intuitively interpretable results that can be displayed as dendrograms or effect plots. CLogitForest is a black-box model with a very strong predictive performance, using variable importance plots for interpretation.

Discussion

The proposed methods are flexible alternatives in the toolbox for analysing matched case-control studies. They can be used both as an exploratory tool for detecting risk and protective factors and as an explanatory tool to provide less biased estimates of a distinct exposure effect.

References

[1] Gunther Schauberger, Luana Fiengo Tanaka, and Moritz Berger (2023). A tree-based modeling approach for matched case-control studies. Statistics in Medicine, 42(5):676-692.

[2] Gunther Schauberger, Stefanie J Klug, and Moritz Berger (2024). Random forests for the analysis of matched case-control studies. BMC Bioinformatics, 25(1):253.

Who Benefits from Intensive Treatment in Psoriatic Arthritis? Deriving Personalised Treatment Rules using Causal Forests and Policy Trees

13:45-14:00

Presenter: Jose Manuel Lema Martinez

Authors: Jose Manuel Lema Martinez [1]; Yunxuan Jiang [1]; Ji-Hyoun Kang [1]; Gabriele De Marco [2]; A.A. den Broeder [3]; Noortje van Herwaarden [4]; Gonul Hazal Koc [5]; Jolanda J. Luime [5]; Helena Marzo-Ortega [2]; Michelle L.M. Mulder [3]; Ilja Tchetverikov [6]; Sofia Massa [1]; Laura Coates [1]; Marc R. Kok [4]; Marijn Vis [7]

Affiliations: [1] University of Oxford, Oxford; [2] University of Leeds, Leeds; [3] Sint Maartensklinieк, Department of Rheumatology, Nijmegen; [4] Maasstad Hospital, Department of Rheumatology and Clinical Immunology, Rotterdam; [5] Erasmus MC, Department of Rheumatology, Rotterdam; [6] Albert Schweitzer Hospital, Department of Rheumatology, Dordrecht; [7] Department of rheumatology, Erasmus MC, Rotterdam

Introduction

Randomised trials of intensive treatment strategies in early, DMARD-naïve psoriatic arthritis (PsA) have produced conflicting results, suggesting clinically meaningful treatment effect heterogeneity obscured when reporting only the average treatment effect (ATE). Causal forests and policy trees allow the estimation of individual-level conditional average treatment effects (CATEs) and translation into interpretable treatment rules.

Methods

We conducted an IPD meta-analysis of four strategy trials (N=349 complete cases N=379 with multiple imputation). Participants received intensive treatment or standard step-up care following EULAR recommendations. The primary outcome was PASDAS, a composite score which combines both clinical outcomes and patient self-assessments, at weeks 16-24 (lower scores indicate better disease control).

Treatment heterogeneity was estimated using causal forests. Robustness was assessed across >40 sensitivity specifications varying propensity score, honest splitting, covariate set, and pooling strategy. Covariates included age, sex, BMI, baseline PASDAS, polyarticular disease, high HAQ, and high CRP. Model performance was assessed via calibration and variable importance.

Personalised treatment rules were derived using policy trees trained on doubly robust pseudo-outcomes with leave-one-study-out cross-fitting. Depth-2 and depth-3 policies were evaluated against treat-all and treat-none benchmarks.

Results

The ATE consistently favoured intensive care across all specifications (range -0.27 to -0.62 PASDAS units). Approximately 27% of participants had CATEs crossing zero. Calibration was satisfactory honest models produced more consistent heterogeneity estimates across specifications. Age and BMI were the main drivers of treatment effect heterogeneity.

Depth-2 trees consistently split first on age then BMI, identifying four subgroups. The primary depth-2 policy improved outcomes over treat-all intensively by 0.16 PASDAS units (range across specifications: 0.11-0.26). Depth-3 trees split first on baseline PASDAS with additional splits on sex and disease duration, yielding larger gains (primary estimate 0.27 range 0.24-0.42). Patients recommended intensive treatment had strongly negative mean effects (-1.37 to -0.73) those recommended standard care had positive estimates.

Discussion

Substantial heterogeneity exists in response to intensive treatment in early PsA. Policy trees derived from causal forests provide interpretable decision rules outperforming uniform strategies. Limitations include modest pooled sample size, heterogeneity in treatment approaches across trials, and absence of external validation.

ZONAR: A Probabilistic Framework for Liver Zonation Mapping Integrating Spatial Omics and Anatomical Regularization

14:15-14:30

Presenter: Michael Debebe

Authors: Michael Debebe [1]; Jörg Rahnenführer [2]; Kathrin Möllenhoff [1]; Achim Tresch [1]

Affiliations: [1] Institute of Medical Statistics and Computational Biology, Faculty of Medicine, University of Cologne, Cologne; [2] Department of Statistics, TU Dortmund University, Dortmund

Introduction

Liver zonation is determined by the intrinsic distances of hepatocytes to the nearest portal and central veins. Inferring these three-dimensional distances from two-dimensional tissue sections is heavily biased by unobserved out-of-plane vessels. Conversely, purely data-driven omics methods are

vulnerable to batch effects and circular reasoning. We introduce ZONAR (ZONation via Anatomical Regularization), a probabilistic framework that overcomes these limitations by combining geometric landmarks with spatial omics data to produce anatomically anchored zonation estimates, which allows us to detect liver metabolic shifts in cholestasis and its treatment from high-resolution imaging mass spectrometry data.

Methods

ZONAR operates on a Delaunay graph of hepatocytes and estimates the zonation coordinate via Maximum A Posteriori (MAP) estimation. The framework combines three components: (i) a molecular likelihood derived from dimension-reduced MALDI-MSI meta-features, modelled as smooth functions of the portal-central position using penalized splines (ii) a geometric-based Gaussian observation model that quantifies local trust in distance-derived zonation ratios and (iii) a Gaussian Markov Random Field smoothness prior that enforces spatial coherence across the tissue. Hyperparameters controlling geometric rigidity and molecular influence are tuned sequentially via connector-level cross-validation.

Results

Applying ZONAR to mouse liver sections analyzed by MALDI-MSI showed that it successfully anchors the zonation field and corrects geometric drift caused by missing landmarks. ZONAR effectively quantified the degradation and blurring of normal metabolic zonation profiles associated with chronic cholestatic injury.

Discussion

Separating spatial priorities from molecular probabilities, ZONAR establishes a statistically sound, anatomically regularized zonation axis. By addressing the individual biases of the geometric and molecular modalities, this framework provides a solid basis for cross-sample measurement of metabolic gradients in spatial omics studies, which is especially helpful for tracking disease progression and treatment effectiveness in pathological conditions such as cholestasis.

Separating What We Cannot Know from What We Do Not Know: Evidential Label Distributions for Nanopore Peptide Event Classification

14:30-14:45

Presenter: Göran Köber

Authors: Göran Köber [1]; Tobias Ensslen [2]; Oliver Amft [1, 3]

Affiliations: [1] Intelligent Embedded Systems Lab, University of Freiburg, Freiburg im Breisgau; [2] Hahn-Schickard, Freiburg im Breisgau; [3] Hahn-Schickard, Villingen-Schwenningen

Introduction

Supervised classification of nanopore translocation events faces a fundamental labeling dilemma: ground truth requires isolating each peptide species individually, which is experimentally challenging at scale, while practical labeling from mixture blockade current histograms introduces structured ambiguity - the partial overlap of ionic current signatures across species - that is physically determined and irreducible at fixed experimental conditions. Existing approaches acknowledge this ambiguity qualitatively but provide no framework for decomposing constituent sources or propagating its structure into the learning objective.

Methods

We introduce an uncertainty-decomposed labeling framework grounded in the Dirichlet-Categorical conjugate model. A variational Bayesian Gaussian mixture model fitted to blockade current histograms yields both soft event-to-species responsibilities and posterior uncertainty over component parameters. The entropy of the responsibility vector quantifies aleatoric uncertainty from irreducible signal overlap parameter posterior width captures epistemic uncertainty from sparse or unbalanced species observations. Both components jointly determine per-event Dirichlet concentration parameters, so that ambiguous events receive diffuse training targets while well-supported events receive sharp ones. An explicit unknown class absorbs out-of-distribution events - artifacts or uncharacterised species - without requiring a separate detector. A state-of-the-art deep neural classifier is trained end-to-end with an evidential loss that penalises overconfidence on low-evidence events.

Results

Applied to nanopore ladder data comprising peptides of varying chain length, labeling from mixture experiments considerably increases usable training examples over single-species collection. The framework assigns lower evidence strength to events near inter-species histogram boundaries and to sparsely observed species. Restricting predictions to high-confidence events yields classification accuracy that exceeds both models trained on conventional FWHM hard labels and uniform soft-label baselines, with boundary species additionally identified as candidates for experimental redesign.

Discussion

The separation of uncertainty sources carries direct operational consequences: high aleatoric uncertainty signals a fundamental measurement limit - no additional data will resolve the ambiguity experimental redesign is indicated. High epistemic uncertainty instead identifies species for which targeted data collection improves label reliability. For diagnostics, calibrated uncertainty scores support reliable deployment decisions, and unknown-class detection flags artifacts automatically. These outputs provide experimentalists a principled, quantitative basis for iterative protocol refinement in nanopore-based diagnostic development.

Reliability of uncertainty estimates under domain shift in time series models

14:45-15:00

Presenter: Sebastian Boie

Authors: Nicolas Frey [1]; Stefan Behling [2]; Niklas Giesa [1]; Matthias Schulte-Althoff [1]; Gregor Lichtner [3]; Elias Grünewald [2]; Felix Balzer [2]; Sebastian Boie [2]

Affiliations: [1] Charité, Institute of Medical informatics, Berlin; [2] Charité - Universitätsmedizin Berlin, Berlin; [3] Universitätsmedizin Greifswald, Klinik für Anästhesie, Intensiv-, Notfall- und Schmerzmedizin, Greifswald

Introduction

Accurate prediction of clinical outcomes in the intensive care unit (ICU) is important for clinical decision support. To achieve high performance, modern prediction models increasingly use high-resolution clinical time series to capture the temporal dynamics of critically ill patients. Beyond predictive performance, probabilistic calibration and uncertainty quantification are crucial for assessing whether model outputs can be trusted in practice, particularly under domain shift. However, models are often trained on data from a single hospital, and their reliability may change when applied to other institutions and diverse populations. We investigate whether uncertainty estimates remain reliable when models are transferred between hospitals on the in-hospital mortality prediction task.

Methods

We train several deep learning architectures for ICU time series modeling, including Long Short-Term Memory networks (LSTM), Gated Recurrent Units (GRU), and Transformer-based models, using the MIMIC-IV database. Predictive uncertainty is estimated using Monte Carlo (MC) dropout, performing multiple stochastic forward passes during inference to obtain a distribution of predictions for each patient. Additionally, we evaluate probabilistic calibration using standard metrics (e.g., expected calibration error) and visualizing calibration curves.

To assess model transferability, models trained on MIMIC-IV are applied to the independent eICU dataset without additional fine-tuning. We evaluate predictive performance on in-hospital mortality, uncertainty estimates, and calibration.

Results

As expected, predictive performance and calibration decreased substantially when transferring models to the independent dataset, indicating dataset shift between institutions. However, uncertainty estimates and calibration metrics did not reliably reflect this performance degradation. In particular, MC dropout uncertainty did not consistently increase for out-of-distribution samples.

Discussion

Presented findings suggest that commonly used uncertainty quantification techniques may be insufficient for detecting performance degradation under domain shift in ICU prediction models. If predictive uncertainty does not increase under performance degradation due to domain shift, clinicians may fail to recognize when a model becomes unreliable. Future work should investigate more

robust methods for domain shift detection and uncertainty quantification in clinical time series models.

Tailoring Statistical Models

Tue, 29.09.26, 13:30–15:00, Room: K3

A typology of methodological research: a consensus study on study types and boundaries

13:30-13:45

Presenter: Stefan Schandelmaier

Authors: Stefan Schandelmaier [1]; Malena Chiaborelli [2]; Julian Hirt [3]; Matthias Briel [4]; Livia Puljak [5]; Zachary Munn [6]; Hanan Khalil [7]; Dawid Pieper [8]; Anne-Laure Boulesteix [9]; Georg Heinze [10]; Richard Hooper [11]

Affiliations: [1] University of Basel, Lörrach; [2] University Hospital Basel, Basel; [3] Department of Clinical Research, Faculty of Medicine, University of Bern, Bern; [4] University of Basel, Basel; [5] University of Split, Split; [6] Adelaide University, Adelaide; [7] La Trobe University, Melbourne; [8] Medizinische Hochschule Brandenburg Theodor Fontane, Rüdersdorf bei Berlin; [9] LMU Munich, Munich; [10] Medical University of Vienna, Wien; [11] Queen Mary University of London, London

Introduction

Methodological research (MR) aims to develop, evaluate, refine, guide, or (de-) implement research methods, or synthesize existing MR. Various types of MR appear in health research journals, but their definitions and terminology remain mostly unclear. This lack of clarity complicates the planning of MR, hinders standard setting, creates uncertainty in labeling, limits indexing in journals and databases, and obscures the concept of methodological evidence. We aimed to develop an initial version of a living, multi-purpose typology of MR.

Methods

An international team of 11 methodologists developed an initial version of the typology through multiple rounds of discussion, feedback, and testing by applying it to MR articles sampled from 26 journals focused on statistical and other methods used in health-related. For each type of MR, we provide a definition, examples, and possible labels. We recorded consensus and remaining disagreements.

Results

We organized types of MR along two cross-cutting dimensions: approach (”how?”, currently 15 types) and purpose (”why”, currently 13 types). Approach-based types include theory-based MR, simulation studies, experimental analyses of real datasets, randomized methodological studies, reliability studies, qualitative studies involving researchers and other interest-holders, consensus studies, reviews of applied methods, meta-epidemiological studies, syntheses of MR. Purpose-based types include proposing new methods comparing methods developing guidance and tools for applying, reporting, or assessing methods and studying methods implementation. The testing revealed challenging types and grey areas related to current reporting of MR. At the conference, we will present

the latest version of the typology (Schandelmaier et al. 2025) and discuss possible applications and research needs.

Discussion

The proposed typology provides a structured framework for classifying MR, helping to clarify terminology and highlight the diversity of approaches and purposes in this field. By making definitions and boundaries explicit, it supports more consistent communication and understanding of MR. Further user-testing is needed to refine and validate its categories.

Reference: Schandelmaier S, Chiaborelli M, Boulesteix AL, Briel M, Pieper D, Hirt J, Heinze G, Puljak L, Mbuagbaw L, Munn Z, Khalil H, Hooper R. Developing a typology of methodological research (TYPE-ME project) [Internet, 2026]. Available from: https://osf.io/4fxaz

---

Formally assessing the maturity of biostatistical methods using the phases of methodological research: an update

13:45-14:00

Presenter: Edouard Bonneville

Authors: Edouard Bonneville [1, 2]; Anne-Laure Boulesteix [3, 2]; Taya Collyer [4, 5]; Alireza Ghorbani [1]; Georg Heinze [6]; Michael Kammer [6, 7]; Tim Morris [8]; Ian White [9]

Affiliations: [1] Institute for Medical Information Processing, Biometry, and Epidemiology, Faculty of Medicine, LMU Munich, Munich; [2] Munich Center for Machine Learning, Munich; [3] Institute for Medical Information Processing, Biometry and Epidemiology, Faculty of Medicine, LMU Munich, Munich; [4] National Centre for Healthy Ageing, Monash University, School of Translational Medicine, Melbourne; [5] Peninsula Clinical School, Monash University, Melbourne; [6] Institute of Clinical Biometrics, Center for Medical Data Science, Medical University of Vienna, Vienna; [7] Department of Medicine III, Division of Nephrology, Medical University of Vienna, Vienna; [8] Statistical Methodology, Novartis Pharmaceuticals UK Ltd., London; [9] MRC Clinical Trials Unit at UCL, University College London, London

Introduction

In biostatistical research, the concept of phases in methodological development, analogous to drug development phases, has been proposed to categorize the maturity and reliability of methods. Heinze et al. (2024, doi: 10.1002/bimj.202200222) identified four phases, ranging from initial method development to comprehensive evaluations and evidence-synthesizing work. The current project aims to refine this concept, providing a framework for identifying both the contribution of a given paper to a method's phase, and the phases that a method has 'completed' based on accumulated evidence.

Methods

First, we reviewed the literature citing the initial work by Heinze et al. to better understand how these methodological phases have been used so far. Second, we developed a tool and checklist designed to extract methodological contributions from literature and map them onto the developmental phases of a method. This tool seeks to facilitate the assignment of cumulative 'badges' that reflect a method's maturity.

Results

The review suggests that the phases concept has often been used as means of 'self-positioning' (e.g. identifying one's study as early phase) for methodological articles, but also for metascientific studies to call for specific kinds of research. We developed a prototype of a tool using guiding questions that address different domains relevant for the assignment of phases. These consider various study-level elements such as mathematical development, simulation studies, case studies, and also method-level characteristics such as robustness to misspecification. Initial experiments suggest that the tool can provide clear signals about a method's phase, helping researchers understand its reliability and applicability for a given task.

Discussion

The proposed framework has implications for various stakeholders, including methodological researchers, method-users, journal editors, and regulators. By providing an indication of a method's maturity, the tool can support method-users in selecting the most appropriate method for their research needs, and highlight areas where further research is needed. Understanding the cultural and psychological factors that underpin the use of a method is crucial for uptake of (what statisticians consider) the most trustworthy methods. A planned Delphi survey will be instrumental in refining the tool and ensuring its applicability across different methodological communities.

Score Plots - Model Diagnostics Made Visual

14:00-14:15

Presenter/Author: Sebastian Büscher

Affiliations: Bielefeld University, Bielefeld

Introduction

Statistical models constitute the backbone of empirical research, enabling estimation of effect sizes with uncertainty quantification, while adjusting for confounding variables.

For sake of interpretability and feasibility, covariate effects are frequently specified in a strictly linear and additive manner, omitting non-linear transformations and interactions.

Such restrictions impose severe assumptions on the estimated model and can lead to model misspecification and biased inference when the data-generating mechanism is more complex.

However, allowing for more flexible model structures can enlarge the search space exponentially, rendering the model selection process intractable.

Despite the use of residual diagnostics in linear models, generally applicable graphical diagnostics are less established for models with binary, ordinal, or categorical outcomes.

Methods

We present score plots, a general graphical model diagnostic tool, for models estimated via optimisation of an additive objective function, including maximum likelihood estimation.

For each observation, the individual score contribution is computed, i.e. the gradient of the objective function with respect to each estimated model parameter, yielding an observation-level score vector.

Plotting score components against covariates provides a direct generalisation of residual plots, as in linear models, the score with respect to the intercept equals the residual.

Examining the dependency of the intercept-related score components on covariate values can indicate functional-form misspecification of the covariates.

Concurrently, patterns in score components of different covariates may suggest omitted interactions.

Results

Across synthetic data scenarios reflecting common departures from additivity (e.g. threshold effects, smooth non-linearity, and interactions), underspecified fitted models produced structured, non-random, and interpretable score plot patterns that correctly flagged the nature of the misspecification.

In applications to real-world datasets, including an investigation of taste and smell loss in type two diabetes patients, score plots revealed plausible non-linear effects and interactions that were not apparent from standard summaries. Subsequent model refinements improved goodness-of-fit significantly.

Discussion

Score plots provide a broadly applicable graphical diagnostic tool for detecting non-linear covariate effects and interactions in likelihood-based (and related) models, including those with non-Gaussian outcomes.

By offering a visual guide for model specification, they can reduce reliance on ad hoc trial-and-error, enhance transparency in model development, and support more reproducible analyses in biostatistical practice.

Regression Trees and Ensemble Methods for Multivariate Outcomes in Clinical and Population Cohorts

14:15-14:30

Presenter: Mousumi Banerjee

Authors: Mousumi Banerjee [1]; Evan Reynolds [2]; Luisa Foco [3]; Martin Gögele [3]; Cristian Pattaro [3]

Affiliations: [1] University of Michigan, Ann Arbor; [2] Michigan State University, East Lansing; [3] Eurac Research, Bolzano

Introduction

Tree-based methods are popular in biomedical research because they offer a flexible, intuitive, and powerful analytic framework for exploring complex data structures. They are often used to predict dichotomous outcomes (e.g. hypertension, neuropathy) even when the outcomes are manifestations of underlying multivariate quantitative measurements (e.g. diastolic and systolic blood pressure nerve conduction measures). Furthermore, existing tree-based methods do not appropriately take into account the correlation that exists in such multivariate data.

Methods

In this paper, we develop goodness-of-split measures for building multivariate regression trees for continuous multivariate outcomes. We propose two general approaches: minimizing within-node homogeneity and maximizing between-node separation. Within-node homogeneity is measured using the average Mahalanobis distance and the determinant of the variance-covariance matrix. Between-node separation is measured using the Mahalanobis distance, Euclidean distance and standardized Euclidean distance. To enhance prediction accuracy we extend single multivariate regression tree to an ensemble of multivariate trees.

Results

Extensive simulations are presented to examine the properties of our goodness-of-split measures. The proposed methods are applied to predict peripheral neuropathy in diabetes patients.

Discussion

Finally, we discuss extensions of our methodology to a causal prediction framework using population data from the Cooperative Health Research In South Tyrol (CHRIS) study for precision prevention of hypertension.

Multi-Study Causal Forest (MCF): Leveraging information from auxiliary data to enhance the estimation of heterogeneous treatment effects

14:30-14:45

Presenter: Ashwini Venkatasubramaniam

Authors: Ashwini Venkatasubramaniam [1]; Julian Wolfson [2]

Affiliations: [1] GSK, Stevenage; [2] University of Minnesota, Minneapolis

Introduction

Precision medicine strategies rely on the ability to identify patient subpopulations who experience greater benefit from treatment. In the past decade, flexible nonparametric methods have been developed using causal inference and machine learning for assessing treatment effect heterogeneity (TEH) within a single study. While auxiliary data may be available, unweighted pooling of such data may lead to biasedness in estimation of conditional average treatment effects (CATEs) for an identified primary source. To overcome this limitation, there is a critical need for methodologies that can effectively leverage "similar" information from auxiliary studies, to enhance the accuracy of CATE estimates. The Multi-Study Causal Forest (MCF) addresses this challenge by extending the known Causal Forest (CF) framework to account for diverse sources of TEH.

Methods

The MCF is a flexible machine learning framework designed to borrow individual patient-level data from auxiliary sources to improve CATE estimation in a primary study of interest. As a novel extension of the CF, the MCF framework is adapted to handle both within-study and between-study heterogeneity, which can arise from varying treatment allocation models, patient populations, or study designs. Our data-driven method leverages the strengths of CFs for capturing complex, nonlinear causal effects to assess the degree of concordance in the CATE functional form across studies, thereby determining the extent of borrowing.

Results

MCF provides robust estimates of TEH and offers improved accuracy over single-study approaches. A comprehensive simulation study demonstrates this superiority of the MCF over a range of scenarios. The MCF method is also illustrated through an empirical application to Breast Cancer data (with primary RCT and auxiliary observational data), to highlight its ability to account for nuanced patterns of effect variation across two distinct study settings. The MCF also compares favourably to an existing methodology, the Multi-Study R-Learner.

Discussion

The MCF framework offers a more accurate and nuanced understanding of how treatment effects vary across individuals and diverse study contexts. By adaptively borrowing information, the MCF significantly enhances the statistical power and interpretability of the treatment effect estimates, especially when single studies are underpowered to detect subtle TEH.

Representation learning for distinguishing case-mix from contextual heterogeneity in model synthesis

14:45-15:00

Presenter: Max Behrens

Authors: Max Behrens [1]; Janis Nolde [2]; Eleni Papakonstantinou [3]; Gabriele Bellerino [4]; Theodoros Evrenoglou [5]; Angelika Rohde [4]; Daiana Stolz [3]; Moritz Hess [6]; Harald Binder [6]

Affiliations: [1] Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, Freiburg; [2] Department of Nephrology, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg im Breisgau; [3] Clinic of Pneumology, Medical Center - University of Freiburg, Freiburg im Breisgau; [4] Department of Mathematical Stochastics, University of Freiburg, Freiburg im Breisgau; [5] Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center-University of Freiburg, Freiburg; [6] Institut für medizinische Biometrie und Statistik, Universität Freiburg, Freiburg

Introduction

When combining regression models across cohorts, whether from separate studies, centers, or sites, standard methods quantify between-cohort heterogeneity (τ^2) but cannot explain its source. Heterogeneity may reflect case-mix differences, meaning cohorts enroll different patient populations but the underlying prognostic relationship is consistent. Or it may reflect contextual differences, meaning similar patients show different prognostic associations depending on cohort due to factors like measurement protocols or care practices. Case-mix heterogeneity can in principle be resolved through standardization, while contextual heterogeneity indicates that a single pooled model may be misleading. Meta-regression attributes τ^2 to known cohort-level moderators, but when relevant moderators are unmeasured or the covariate space is high-dimensional, the question of whether to pool remains unresolved.

Methods

We fit cohort-specific local regression models for a continuous outcome across the covariate space to capture how prognostic relationships vary within and between cohorts. To make local estimation feasible with many predictors, we learn a low-dimensional representation via an autoencoder jointly optimized with a composite loss that balances reconstruction, local prognostic fit, and cohort discrimination, so that patient proximity reflects similarity in outcome relationships. A variance decomposition then partitions total prognostic variance into between-cohort components (case-mix, context, and their covariation) and residual within-cohort variation. We illustrate this on a two-center COPD dataset (318 patients, 36 lung function and physiological predictors) and assessed the decomposition through inverse propensity weighting (IPW) and perturbation analyses.

Results

Context dominated in the between-cohort variance as 38% was contextual and 29% reflected covariation, while pure case-mix contributed only 13%. Consistent with contextual dominance, IPW reduced between-cohort heterogeneity by only 6.4%. Perturbation analyses showed that applying the wrong cohort's local model increased mean squared prediction error by 90%, compared to 52%

when applying models to mismatched patient distributions. Results were stable across 10 independent random seeds.

Discussion

The approach complements standard multi-cohort regression synthesis by attributing heterogeneity to interpretable sources, which may help inform pooling. When case-mix dominates, reweighting may support pooling. When context dominates, as in our illustrative application, pooling may not be appropriate. Unlike meta-regression, the decomposition does not require pre-specified moderators and accommodates high-dimensional predictor spaces.

Strategies for Evidence Synthesis

Tue, 29.09.26, 13:30–15:00, Room: K4

Chair: Adriani Nikolakopoulou

Systematic Reviews and Meta-analysis of Prognosis Research and Prediction Models

13:30-13:45

Presenter: Richard Riley

Authors: Richard Riley [1]; Karel Moons [2]

Affiliations: [1] University of Birmingham, Birmingham; [2] UMC Utrecht, Utrecht

Introduction

Understanding, summarising and predicting outcomes is essential to modern healthcare planning and individual decision-making. This has led to a rapid rise in prognosis research studies, including thousands of studies developing and evaluating prediction models (algorithms) for individual outcome prediction. However, primary prognosis studies have highly variable quality and often contain conflicting findings. This motivates the need for systematic reviews of prognosis research and prediction model studies, including meta-analysis methods to summarise evidence about, for example, model performance and prognostic factor effects. The aim of this talk is to outline current methods and challenges for conducting systematic reviews of prognosis research and prediction models, including meta-analysis, and to highlight the work and recommendations of the Cochrane Prognosis Methods Group.

Methods

The PROGRESS framework is used to define three types of prognosis review, focusing on: (i) average prognosis of individuals (overall prognosis reviews), (ii) factors associated with a particular outcome (prognostic factor reviews), and (iii) multivariable models for individual outcome prediction (prediction model reviews). Review methods for each type are outlined, and the talk covers the steps required throughout the whole review process, using real examples for illustration, as recommended by the Cochrane Prognosis Methods Group.

Results

Reviews should be framed using PICOTS (Population, Index, Comparator, Outcomes, Timing and Setting), adapted for each review type. Study searching and screening is time-consuming, due to a lack of taxonomy in the prognosis literature. For included studies, recommended tools for assessing risk of bias and applicability are ROB-OP (overall prognosis), ROB-PF (prognostic factors) and PROBAST+AI (prediction models), with data extraction following the CHARMS guidelines. Indirect estimation of study results may be required, especially for prognostic factor effects and prediction model performance. Meta-analysis may not always be feasible or sensible if done, heterogeneity is

anticipated and so random-effects models and prediction intervals are recommended. Small-study effects should be examined, and TRIPOD extensions followed when reporting a review.

Discussion

Prognosis reviews are increasingly demanded but require specific methods tailored to overall prognosis, prognostic factor and prediction model review questions, including bespoke methods for assessing risk of bias, extracting statistical estimates and undertaking meta-analysis.

Site-stratified posterior divergence from Prior-Fitted Networks as an external validity diagnostic for multi-site clinical prediction

13:45-14:00

Presenter: Patric Tippmann

Authors: Patric Tippmann; Max Behrens; Harald Binder

Affiliations: University of Freiburg, Freiburg, Germany, Freiburg

Introduction

Prior-Fitted Networks (PFNs) achieve competitive predictive accuracy on small-to-medium clinical tabular datasets through amortized Bayesian inference. Their approximate posteriors provide uncertainty estimates that could inform external validation, yet whether site-stratified posterior divergence can serve as an external validity diagnostic is unexamined. The standard heterogeneity measure τ^2 quantifies between-study coefficient variation but does not directly predict site-specific calibration failure, while covariate-only distributional comparison misses outcome-relevant heterogeneity.

Methods

We propose an external validity diagnostic based on site-stratified PFN posterior divergence. We condition a PFN on each site's data independently and obtain predictive distributions on a shared evaluation set. Divergence between each site's posterior and the pooled posterior yields a per-site heterogeneity score. We evaluate whether this score predicts held-out calibration error from leave-one-site-out cross-validation on a multi-center clinical dataset (International Stroke Trial, 24 sites, N=18,724) with between-site differences in covariate distributions and outcome prevalence. We further investigate the divergence at the feature level to identify which predictors are most associated with calibration failure at external sites, and compare against a covariate-only distributional measure.

Results

PFN posterior divergence strongly predicted held-out calibration error across 24 sites (Spearman ρ = 0.93, $p < 0.001$). The association remained strong after controlling for outcome prevalence differences (partial ρ = 0.83, $p < 0.001$), confirming that PFN divergence captures heterogeneity beyond

simple prevalence mismatch. A covariate-only distributional measure showed substantially weaker association with calibration error (Mahalanobis distance ρ = 0.22, p = 0.31). Per-feature decomposition of the divergence identified stroke subtype classification as most strongly associated with between-site heterogeneity, consistent with known differences in subtype classification practices across participating centers.

Discussion

Site-stratified PFN posterior divergence provides a practical external validity diagnostic that substantially outperforms covariate-only distributional comparison. Conditioning on the outcome - rather than comparing covariate distributions alone - is consistent with this finding. Per-feature decomposition can identify predictors most associated with between-site miscalibration, enabling site-specific assessment of which model components may require recalibration before external deployment.

Building a decentralized pipeline for two-stage IPD meta-analysis: methodological and practical lessons from SPEED MS

14:00-14:15

Presenter: Marianne Charlotte Tokic

Authors: Marianne Charlotte Tokic [1]; Romy Lauer [2]; Sandra Thiel [3]; Sabrina Haben [4]; Kerstin Hellwig [5]; Nina Timmesfeld [6]

Affiliations: [1] Ruhr University Bochum, Department of Medical Informatics, Biometry and Epidemiology, Bochum; [2] Ruhr-Universität Bochum, Abteilung für Medizinische Informatik, Biometrie und Epidemiologie, Bochum; [3] Department of Neurology, St. Josef-Hospital, Ruhr-University Bochum, Bochum; [4] Ruhr University Bochum, Department of Neurology, St. Josef Hospital, Bochum, Germany, Bochum; [5] Ruhr-Universität Bochum, Universitätsklinik für Neurologie, St. Josef Hospital, Bochum; [6] Abteilung für Medizinische Informatik, Biometrie und EpidemiologieRuhr-Universität Bochum, Bochum

Introduction

Two-stage individual participant data meta-analysis (IPD-MA) is a flexible approach for collaborative evidence synthesis, especially when individual-level data sharing is restricted. However, implementation is often challenged by heterogeneous source data, variable governance requirements, and the need for reproducible analytical workflows. We developed a decentralized pipeline for two-stage IPD meta-analysis and present key methodological and practical lessons from its use in SPEED MS, an ongoing IPD network meta-analysis in multiple sclerosis.

Methods

The pipeline combines three core elements: a pre-defined cross-study variable table, reusable standardized analysis scripts, and a twostage synthesis framework. Participating study centers may either transfer harmonized IPD for centralized analysis or run the predefined scripts locally and share aggregated study-specific estimates and corresponding uncertainty measures only. This design allows consistent first-stage analyses despite heterogeneous data-sharing conditions. SPEED MS serves as a case example for estimating sex-specific treatment effects across randomized trials of disease-modifying therapies in relapsing-remitting multiple sclerosis.

Results

The pipeline improved early collaboration with study centers by clarifying variable requirements prior to data transfer and by supporting transparent harmonization procedures. Standardized scripts reduced analytic variability across centers and enabled governance-compatible decentralized first-stage analyses. Key lessons learned include the importance of a priori variable specification, the need to balance methodological rigor with realistic data availability, and the value of script-based standardization for reproducibility and feasibility in collaborative IPD projects.

Discussion

A decentralized two-stage IPD-MA pipeline can improve transparency, reproducibility, and feasibility in collaborative meta-research. Our experience suggests that pre-specification of variables, reusable scripts, and flexible governance-compatible workflows are central to successful implementation, particularly in complex settings such as IPD network meta-analysis. This framework may serve as a practical blueprint for future decentralized IPD-MAs and IPD-NMAs.

Small-sample hypothesis testing for binomial meta-analyses

14:15-14:30

Presenter: Riccardo De Santis

Authors: Riccardo De Santis [1]; Ruggero Bellio [2]; Annamaria Guolo [1]

Affiliations: [1] University of Padova, Department of Statistical Sciences, Padova; [2] University of Udine, Italy, Udine

Introduction

Binomial outcomes arise frequently in medical studies, where information can be summarized using sensitivity and specificity. When studies are aggregated in a meta-analysis, the most popular inferential approach relies on the normal approximation of the logit-transformed sensitivity and specificity. However, this approximation can be inaccurate for studies with small-to-moderate sample sizes for instance, it can lead to inflated type I error rates in hypothesis testing. Alternatively, accounting for the binomial nature of the data requires a binomial random effects model, whose estimation may be computationally intensive and exhibit poor small-sample performance.

Methods

To address the above-mentioned issues, we propose adopting the flip-score approach for hypothesis testing, a semiparametric method previously suggested for inference in generalized linear models. The key idea relies on the fact that we do not rely on the usual normal – or Student's t – approximation of some test statistics, but we tackle the problem with the use of a discrete distribution using the sign-flips, which are naturally close to the more known permutation-based tests. Our proposal does not come only with simulation-based evidence, but also with theoretical guarantees.

Results

Simulation studies demonstrate that this method improves type I error control in small samples while maintaining competitive power relative to standard approaches. We evaluate the performances under different scenarios, based on usual binomial meta-regression settings: we also discuss two different ways of specifying the random effects, when generating synthetic data, and their practical implications.

Discussion

Random effects can be problematic to handle in case of non-normal outcomes, since the induced mixed models are naturally complex by nature. On the other hand, they cannot be simply ignored, as the simpler models would be unreliable. In this sense, our approach aims to provide a more reliable tool for hypothesis testing in terms of type I error control under the null hypothesis, but without a relevant loss of power when such hypothesis is not true.

Sensitivity analysis for publication bias in rare-event meta-analysis under generalized linear mixed models

14:30-14:45

Presenter: Yi Zhou

Authors: Yi Zhou [1]; Taojun Hu [2]; Satoshi Hattori [3, 4]

Affiliations: [1] Division of Mathematics and Informatics, Graduate School of Human Development and Environment, Kobe University, Kobe; [2] School of Medicine, Chongqing University, Chongqing; [3] Department of Biomedical Statistics, Graduate School of Medicine, The University of Osaka, Suita; [4] Integrated Frontier Research for Medical Science Division, Institute for Open and Transdisciplinary Research Initiatives, The University of Osaka, Osaka

Introduction

In systematic reviews and meta-analyses, publication bias (PB) is one of the serious concerns and mainly induced by selective publication of academic literatures. Although many methods have been proposed to address PB, almost all of them are based on the normal-normal (NN) random-effects model, assuming that data are normally distributed at both the within-study and between-study levels. For rare-event meta-analysis where data contain rare occurrences of events, the standard NN

random-effects model may perform poorly. Instead, some generalized linear mixed models (GLMMs) which employ the exact distribution for the number of events at the within-study level increase the estimation accuracy and have been widely used in practice. However, limited methods can be applied to address PB in the GLMMs. To address this limitation, we propose a framework of sensitivity analysis for evaluating the impact of PB in the contrast-based GLMMs.

Methods

The proposed framework is developed based on the famous Copas-Heckman-type sensitivity analysis methods by relating study-specific true effect sizes with the latent Gaussian variable on the study sample sizes for adjusting PB. The conditional likelihood function was derived for estimating the overall effect sizes constrained by probabilities of publications as sensitivity parameters.

Results

Simulation studies are conducted to assess the performance of the proposed methods in adjusting PB and compare the results with related existing methods. When the publication process was correctly specified, the proposed methods gained the least bias. Real-world examples showed the broad applicability of our proposal in evaluating the potential impact of PB in meta-analysis of odds ratios and proportions with rare-event outcomes.

Discussion

In systematic reviews and meta-analyses, PB has been recognized as an inevitable issue. To deal with PB, many advanced selection-model-based methods have been proposed. However, implementing these methods for rare-event meta-analyses can be challenging and problematic since they are developed based on the NN model. The proposed methods serve as useful tools for assessing the robustness of results synthesized by GLMMs, overcoming the lack of methods for addressing PB in rare-event meta-analyses. The proposed methods completely avoid the need for continuity corrections and can be easily implemented using standard software with low computational cost.

Literature-based meta-analysis of adverse events accounting for heterogeneous follow-up duration in oncology clinical trials

14:45-15:00

Presenter/Author: Satoshi Hattori

Affiliations: The University of Osaka, Suita

Introduction

It is difficult to understand the safety profile of drugs based on a single clinical trial since clinical trials are often designed to prove efficacies and sample size is not powered for safety assessment. Thus, meta-analysis would be a valuable tool to infer the safety profiles. Individual clinical trials usually report the incidence proportions of adverse events (AE). The follow-up duration may be study-specific, and furthermore different between the treatment groups within a single study. It often occurs

in oncology clinical trials subjects are usually followed until the primary endpoint such as progression-free survival (PFS) is fixed. If this is the case, it is hard to interpret the aggregated relative risk with the standard meta-analysis techniques.

Methods

The PFS or the overall survival (OS) is often used as the primary endpoint in oncology clinical trials and the Kaplan-Meier estimates of the survival functions for them are often demonstrated graphically to highlight the efficacy results. We propose novel safety meta-analysis methods that address difference in follow-up duration for AEs among the treatment groups. Our idea is to analyze the AE as a time-to-event outcome by considering the cumulative incidence function as the estimand utilizing the Kaplan-Meier estimates of the time-to-event primary endpoint, which form the distribution of the follow-up duration of AEs. At first, we introduce our proposed method assuming that the follow-up duration (e.g. PFS, OS) is independent of the time-to-AE. We also introduce an inference procedure

accounting for dependence between the time-to-AE and the follow-up duration with copulas.

Results

We showed theoretical property of the proposed methods. Our simulation study revealed that our proposed methods worked as designed unless the event rate of AE was not extremely low. We illustrate our proposed method by re-analyzing the meta-analysis of cerebrovascular events of antineoplastic agent bevacizumab.

Discussion

As far as we know, our approach is the first literature-based meta-analysis method for AEs that can address difference in follow-up duration. To compare the risk and benefit of drugs, proper adjustment of the heterogeneous follow-up durations should be considered and the proposed method would be useful in such situations.

Adaptive Monitoring and Borrowing in Trials

Tue, 29.09.26, 13:30–15:00, Room: K5

Chair: Andrea Callegaro

---

e-values instead of p-values in clinical trials: What happens?

13:30-13:45

Presenter: Yongxi Long

Authors: Yongxi Long [1]; Alexander Ly [2]; Nikos Ignatiadis [3]; Jelle Goeman [1]; Peter Grünwald [2]; Erik van Zwet [1]

Affiliations: [1] Leiden University Medical Center, Leiden; [2] Centrum Wiskunde & Informatica, Amsterdam; [3] The University of Chicago, Chicago

Introduction

e-values have been proposed as a flexible alternative to p-values that enable anytime-valid inference. In the context of clinical trials, they allow continuous monitoring of accumulating evidence ("peeking") and optional continuation of recruitment. However, this flexibility comes at the cost of reduced statistical power. Our goal is to assess this trade-off empirically.

Methods

We estimated the distribution of effect sizes from more than 20,000 randomized controlled trials from the Cochrane Database of Systematic Reviews (CDSR)., We then generated synthetic trial trajectories up to and beyond their original sample sizes. As a natural choice, we evaluated the e-process given by the Bayes factor of this CDSR-derived effect size distribution relative to the null model.

The alpha-spending speed of e-processes is tunable to prioritize detecting either small or large effects: large effects can be captured faster by a more aggressive e-process at the price of excessive sample sizes for small/medium effects. Design information including the minimum clinically important difference (MCID) can be incorporated to better reflect the goals and constraints of a particular trial.

Results

The empirical Bayes e-process is highly conservative because it reserves its alpha-budget for detecting many small-effect trials in the CDSR cohort at large sample sizes. Therefore, it requires approximately 3 times the original sample size to match the probability of significance (PoS) of a standard fixed-sample p-value test. In contrast, an MCID-informed e-process requires, on average, about 1.8 times the original sample size, but actually uses only about 1.5 times the original sample size due to optional stopping. Smoothing the MCID-informed e-process using a method-of-moment prior

improves robustness to effect-size misspecification for individual trials, but pays a larger efficiency price on average across the CDSR cohort.

Discussion

The great flexibility of e-values offers attractive advantages, particularly in enabling anytime-valid testing. Their practical implementation in clinical trials requires a balance between flexibility and power. In a setting characterized by many small effects and resource constraints, more aggressive e-process constructions yield better average efficiency.

Adaptive clinical trials based on design-optimal e-values: A novel approach for efficient and n_max anytime-valid inference in single-arm trials with binary data

13:45-14:00

Presenter: Joost van Rosmalen

Authors: Stef Baas [1]; Judith ter Schure [2]; Joost van Rosmalen [3]

Affiliations: [1] MRC Biostatistics Unit, University of Cambridge, Cambridge; [2] Amsterdam UMC, Amsterdam; [3] University Medical Center Utrecht, Utrecht

Introduction

The e-value has recently been gaining attention in the statistical literature as a more robust alternative to p-values and Bayes factors for assessing the strength of evidence against a statistical hypothesis. E-values are a promising methodology for adaptive clinical trials due to their anytime-validity: e-values ensure type I error rate control, at any analysis time, even if this time depends on previously observed data, thereby facilitating repeated interim analyses and complex stopping rules.

Methods

We investigate the performance of e-values in single-arm multi-stage clinical trials with binary data. This setting is relevant in early-phase cancer trials, but it also facilitates an accessible introduction to the betting interpretation, which can be used to construct e-values as well as stopping rules based on e-values. Instead of the commonly proposed growth-rate optimal e-values, we introduce design-optimal e-values that follow from betting strategies that either (1) maximize statistical power at a maximum sample size nmax, or (2) minimize the expected sample size, with or without constraints on the minimum power. To construct design-optimal e-values, we employ an efficient dynamic programming approach based on the currently observed e-value, the maximum sample size, and the pre-specified significance level.

Results

Using exact calculations, we present examples including and excluding optimal futility stopping boundaries based on e-values, and illustrate the operating characteristics. Our evaluation shows that

stopping rules based on design-optimal e-values can yield similar or better operating characteristics than other adaptive and non-adaptive designs, as well as growth-rate-optimal e-values.

Discussion

Design-optimal e-values can provide a viable alternative to the current state-of-the art in single-arm binary trials, and warrant extension to other adaptive clinical trial settings such as multi-arm multi-stage and response-adaptive designs.

A conditional bias adjustment method based on a truncated Jeffreys prior for already-reported clinical trials stopped early for efficacy

14:00-14:15

Presenter: Shoki Izumi

Authors: Shoki Izumi [1, 2]; Shogo Nomura [3]; Mitsunori Ogawa [4]; Yutaka Matsuyama [5]

Affiliations: [1] Department of Biostatistics, Division of Health Sciences and Nursing, Graduate School of Medicine, The University of Tokyo, Tokyo; [2] Section of Biostatistics, Department of Clinical Data Science, Clinical Research & Education Promotion Division, National Center of Neurology and Psychiatry, Tokyo; [3] Department of Biostatistics and Bioinformatics, Graduate School of Medicine, The University of Tokyo, Tokyo; [4] Department of Health Data Science, Tokyo Medical University, Tokyo; [5] Department of Biostatistics, School of Public Health, Graduate School of Medicine, The University of Tokyo, Tokyo

Introduction

Confirmatory clinical trials that terminate early for efficacy often draw considerable attention from medical community, regulators, and public. For reasons that treatment effect estimates after early stopping can be conditionally-biased and the confidence intervals (CI) may fail to maintain nominal conditional coverage, appropriate bias-adjustment methods are required to mitigate the conditional bias in an effect estimate. Current issues surrounding existing bias-adjustment methods however do not satisfy all the following properties: (A1) individual-level patient data is not required (A2) CI never results in an empty set (A3) CI contains the associated point estimate (A4) point estimate is compatible with the test result (B1) CI is not excessively wide (B2) CI is compatible with the test result (B3) CI preserves conditional coverage at nominal level (B4) point estimator does not have conditional bias in the direction of underestimation, which is the opposite direction to that of the unadjusted estimator. Deviation of property A1 is critical because almost all the stakeholders cannot access individual-level patient data in general. The aim of this study is to assess whether a Bayesian approach with a weakly informative prior can address these issues, with primary emphasis on properties A1-A4.

Methods

We propose a novel truncated Jeffreys prior based weakly informative prior. Bayesian estimation procedure can solve the issues properties A1-A3. Truncation is required for satisfying the property A4. Although both maximum a posteriori (MAP) estimator and posterior mean were considered adjusted methods, we primarily focused on the MAP estimator. Properties A1 and A3 were assured theoretically.

Results

Simulation studies indicated that the proposed method satisfies properties A2, A4, B1, B2, and B4, and roughly preserves conditional coverage at the nominal level (cf. property B3). These desirable properties of the proposed method were also observed in applications to actual clinical trials.

Discussion

It was suggested that the proposed method roughly satisfies all of the desirable properties A1-A4 and B1-B4, which no existing method was able to satisfy simultaneously.

On the inclusion of non-concurrent controls in platform trials with an interim analysis

14:15-14:30

Presenter: Pavla Krotka

Authors: Pavla Krotka [1]; Martin Posch [2]; Marta Bofill Roig [1]

Affiliations: [1] Universitat Politècnica de Catalunya, Barcelona; [2] Medical University of Vienna, Vienna

Introduction

Platform trials efficiently evaluate multiple treatments simultaneously, allowing new arms to enter over time while sharing a common control. Interim analyses add further flexibility by enabling early stopping. For late-entering arms, incorporating non-concurrent controls (NCC) into the analysis can reduce the required sample size and increase power. However, it might also introduce bias in the effect estimates and hypothesis tests if time trends are present.

Methods

To evaluate the impact of an interim analysis in platform trials with continuous endpoints utilizing NCC, we study a 2-arm trial with a shared control, where Arm 2 enters later. An interim analysis for Arm 1 is performed at the time Arm 2 is added, allowing early stopping for futility and/or efficacy. For the final analysis of Arm 2, we consider a frequentist regression model that uses NCC and adjusts for time using a categorical covariate. We derive the bias of the resulting estimator of the Arm 2 effect, which arises because the borrowing of NCC now stochastically depends on the interim outcome in Arm 1. We then propose a new mean adjusted estimator (MAE) for Arm 2, which corrects for this bias by subtracting the conditional bias given that Arm 1 continues.

Results

The unadjusted model-based estimator for Arm 2 is biased once interim stopping for Arm 1 is enabled. The magnitude of the bias depends on design parameters that influence the extent of NCC borrowing, i.e., the sample sizes, the interim stopping rule, and the treatment effect in Arm 1. In particular, the marginal bias of the unadjusted estimator is maximized when Arm 1 has a 50% probability of continuing. The newly proposed MAE substantially reduces both marginal and conditional bias and type I error inflation while leading to power gains compared to an analysis using only concurrent controls.

Discussion

Existing methods for NCC inclusion have not accounted for interim analyses. We show that conducting an interim analysis in one arm affects the effect estimation in other arms when NCC are used. Mean adjusted estimators provide a practical approach to restore valid inference while preserving some of the efficiency gains from utilizing NCC.

A Bayesian framework for assurance-based design and borrowing information in clinical trials with application to trials in rare diseases

14:30-14:45

Presenter/Author: Yi Huang

Affiliations: University of Newcastle, Newcastle

Introduction

Small-sample and rare disease trials face substantial uncertainty in treatment effects, making conventional power-based design unreliable. While historical data can be used to improve efficiency, they may also introduce bias if they are not consistent with the current trial. We propose a Bayesian framework that integrates assurance-based design, borrowing of historical information, and operating characteristics to evaluate trial performance in terms of relative assurance, Type I error rates, and decision probabilities such as false positive possibility.

Methods

We develop a Bayesian design for a placebo-controlled trial using an ANCOVA model. Uncertainty in the treatment effect is represented through design priors, while historical control data are incorporated using meta-analytic predictive (MAP) priors and robust extensions (rMAP), which allow information borrowing while protecting against prior-data conflict. Posterior inference is performed using Markov chain Monte Carlo (MCMC), and operating characteristics are evaluated via simulation under the prior predictive distribution.

Results

Borrowing increases assurance and reduces required sample size, indicating improved efficiency. However, when historical data differ from the current trial, borrowing can inflate false positive conclusions. The rMAP approach reduces this risk by down-weighting historical information when inconsistency arises. We show that designs with similar type I error can have different probabilities of success and false positives, highlighting that borrowing changes the overall behaviour of the trial, not just error rates.

Discussion

This work brings together three key components - assurance for trial design, Bayesian borrowing of historical information, and Bayesian operating characteristics within a unified design and analysis framework tailored to small-sample settings such as rare disease trials. By linking these elements, the framework quantifies how borrowing influences both efficiency and the risk of incorrect conclusions. The redesign of existing trials is used as an example to illustrate how this approach can be used in practice to investigate design properties and inform decision-making.

Time-Dependent Bayesian Borrowing from Historical Controls in Randomized Clinical Trials with Temporal Drift

14:45-15:00

Presenter: Emmanuel Adewuyi

Authors: Emmanuel Adewuyi [1]; Jonathan Bartlett [1]; Darren Scott [2]; Alex Lewin [1]

Affiliations: [1] London School of Hygiene & Tropical Medicine, London; [2] AstraZeneca, Cambridge

Introduction

Borrowing information from historical controls can improve efficiency in randomized clinical trials however, standard borrowing approaches typically assume exchangeability across studies. This assumption may be inappropriate when standards of care improve over time, inducing temporal drift between historical and current control outcomes. We propose a Bayesian model with a time-dependent informative prior for borrowing from multiple historical studies to improve estimation of the treatment effect in the current trial.

Methods

We consider individual-level continuous outcomes from one current randomized trial and multiple historical control studies. We specify a prognostic score-adjusted outcome model in which prognostic covariate effects are shared across studies, while study-specific control intercepts evolve over calendar time. Temporal dependence among these intercepts is modeled through an autoregressive hierarchical prior, allowing borrowing to be stronger for more commensurate studies and weaker when temporal drift is present. Posterior draws are combined with g-computation to estimate the

current-trial marginal mean under treatment, the marginal mean under control, and their difference. A simulation study compares the proposed method with no borrowing, full pooling, power priors, and exchangeable hierarchical priors under scenarios with no, mild, and strong drift.

Results

The simulation study is designed to assess bias, interval coverage, precision, and the degree of effective borrowing across drift scenarios. We expect the proposed approach to improve precision when drift is absent or mild, while remaining robust in the presence of substantial temporal drift. Final results will quantify efficiency gains and protection against bias. In addition, a real randomized clinical trial dataset will be analyzed to evaluate performance and assess consistency with the simulation findings.

Discussion

Time-dependent informative priors provide a principled framework for adaptive borrowing from historical controls when outcome processes evolve over time. By combining temporal modeling with covariate adjustment and marginal effect estimation, the proposed approach seeks to improve efficiency while preserving validity in contemporary clinical trials.

Target Trial Emulation

Tue, 29.09.26, 13:30–15:00, Room: N1

Chair: Derek Hazard

The ideal trial: defining causal estimands that balance relevance and feasibility in target trial emulations and actual randomized trials

13:30-13:45

Presenter: Margarita Moreno-Betancur

Authors: Margarita Moreno-Betancur; Rushani Wijesuriya; John B. Carlin

Affiliations: Murdoch Children's Research Institute and University of Melbourne, Melbourne

Introduction

Causal inference is the goal of randomized trials and many observational studies. The first step in a formal causal inference framework is to define the causal estimand, and in both types of study this can be done mathematically or, equivalently, by specifying an ideal trial: a hypothetical perfect randomized experiment (with representative sample, perfect adherence, etc.). The target trial framework is increasingly used for causal inference in observational studies, but clarity is lacking in how a target trial should be specified and how it relates to an ideal trial.

Methods

Here, we review the mathematical and ideal trial approaches to defining a causal estimand. We then consider the question of how a target trial should be specified, reviewing a recommended approach, commonly seen in applications: to specify the target trial such that it is closely aligned with the observational data (e.g. uses the same eligibility criteria). An example from respiratory epidemiology is used for illustration of concepts.

Results

We show the equivalence between the mathematical and ideal trial approaches to estimand definition and highlight the need to balance practical relevance and feasibility of estimation regardless of approach. We then outline the challenges of the "aligned" approach, which puts the focus heavily on feasibility of estimation. Specifically, we show that with this approach biases may remain relative to the estimand of ultimate practical interest, defined by the ideal trial, that mirror the often-overlooked biases of actual trials. We further outline how the concepts discussed apply to actual trials.

Discussion

We conclude that consideration of the ideal trial and of how the target trial and its emulation or the actual trial differ from it is necessary to identify and manage all bias sources in both settings.

Target trial emulation within a trial? Per-protocol approaches for built-in treatment switches in trials for ethical reasons

13:45-14:00

Presenter: Rik van Eekelen

Authors: Rik van Eekelen [1]; Saskia Klein Meuleman [2]; Emiel Post Uiterweer [2]; Robert de Leeuw [2]; Judith Huirne [2]

Affiliations: [1] Department of Epidemiology and Data Science, Amsterdam UMC location Vrije Universiteit, Amsterdam; [2] Amsterdam UMC, Amsterdam

Introduction

Recently, we conducted and analysed a randomised controlled trial in which subfertile women with an caesarean scar defect were randomly allocated to scar surgery or no surgery (i.e. expectant management).(1) The endpoint was ongoing pregnancy and participants were followed for two years. For ethical reasons, participants randomised to expectant management were offered to have the surgery after nine months. In addition, participants receiving surgery needed to recover for up to six months before again trying to conceive. This yielded multiple methodological complications in the analysis.

Note that this research is currently under embargo such that no actual clinical results can be shared until this embargo is lifted.

Methods

We approached the problem of built-in treatment switches from several angles. We started with an intention-to-treat approach (ITT, following ICH E9(R1)'s treatment policy), which had inherent limitations to its interpretation. For didactical reasons we continued with a flawed 'classical' per-protocol (PP) analysis (conditional on adherence i.e. no switch for two years). Next, following the literature we approached it as an observational design and applied the target trial emulation (TTE) framework, defining strategies participants needed to adhere to (hypothetical), or using all data including after switching via a time-varying covariate for surgery (as-treated (AT), while on treatment).(2) These latter approaches used inverse probability weighting or g-computation for conditional exchangeability of treatment groups.(2)

Results

We show the study timeline to explain which time periods had clear clinical interpretations using conventional methods and which did not. If possible we show actual results, otherwise simulated examples. The classical per-protocol analysis is clearly erroneous and the ITT does not have a useful interpretation, whereas the TTE hypothetical approach yielded the clearest interpretation with arguably the most clinical utility. All approaches had built-in non-proportionality of hazards. Incorporating data after the cross-overs did not increase precision.

Discussion

We discuss the meaning of ITT, PP and AT in RCTs with built-in treatment switches and argue that terminology from the TTE framework and the ICH E9(R1) addendum helps clarify estimands.

References

1. Trial protocol, https://tinyurl.com/4swd7xsa

2. Hernán MA, Robins JM. Per-Protocol Analyses of Pragmatic Trials. N Engl J Med. 2017 Oct5377(14):1391-1398. doi: 10.1056/NEJMsm1605385.

Time Partitioning in Target Trial Emulation

14:00-14:15

Presenter: Harold Tankpinou Zoumenou

Authors: Harold Tankpinou Zoumenou [1]; Simon Ferreira [1]; Charles Assaad [1]; Nathanaël Lapidus [1]; Daria Bystrova [1]; Benjamin Glemain [1, 2]

Affiliations: [1] Institut Pierre Louis d'Epidémiologie et de Santé Publique, Sorbonne Université, INSERM, Paris; [2] Sorbonne University, Paris Brain Institute (ICM), INSERM, INRIA, CNRS, APHP, Paris

Introduction

In target trial emulation, treatment initiation is often time-dependent and misaligned with the time of inclusion, which may lead to immortal time bias. Cloning-censoring-weighting (CCW), which is commonly used in this setting, relies on partitioning follow-up time. However, depending on the clinical context, the same partition may lead to different causal relationships between treatment and outcome within time intervals. Thus, the validity of CCW may depend on time partitioning. In this work, we use the framework of structural causal models (SCMs) to make explicit the causal assumptions implied by time partitioning and assess the validity of CCW under each causal structure.

Methods

Using SCMs, we formalized two trial emulation scenarios that differed in the lag of the treatment effect, examined the relevance of daily time partitioning, and assessed the identifiability of treatment effect and the validity of CCW using graphical criteria and simulations.

Results

In the first scenario (approximately a one-month lag for the treatment effect), daily partitioning led to an identifiable treatment effect but also to high model dimensionality. This dimensionality could be reduced by widening the time intervals, provided that treatment did not affect the outcome within these intervals. In the second scenario (a lag of a few hours), daily partitioning led to an identifiable treatment effect only if death occurring on a given day did not preclude treatment initiation during that same day (for example, if treatment can only be administered at the very beginning of the interval). Graphical criteria and simulations supported the validity of CCW when treatment did not affect the outcome within time intervals (estimated bias: -0.06%, 95% CI: -0.24% to 0.14%), but not otherwise (estimated bias: -7.94%, 95% CI: -8.13% to -7.76%).

Discussion

In target trial emulation, certain time partitions may result in causal structures that invalidate CCW. More broadly, time partitioning should account for the clinical context, and in particular the time to treatment effect.

Trial emulation to assess the effect of surgery on survival when there are competing risks, with application to patients with thoracic aortic aneurysms

14:15-14:30

Presenter: James Murray

Authors: James Murray [1]; Caroline Chesang [1]; Steve Large [2]; Colin Bicknell [3]; Carol Freeman [2]; Ruth Keogh [1]; Linda Sharples [1]

Affiliations: [1] London School of Hygiene & Tropical Medicine, London; [2] Royal Papworth Hospital NHS Foundation Trust, Cambridge; [3] Department of Surgery and Cancer, Imperial College London, London

Introduction

Randomised controlled trials (RCTs) are the gold standard for causal inference but are often infeasible for surgical interventions due to logistical and ethical complexities. Selection-for-treatment bias arises from surgeon and patient preferences, and immortal time bias occurs because patients must survive until surgery. Failure to account for these biases overestimates surgical benefits.

Methods

We use trial emulation and a cloning-censoring-weighting approach (CCW) to estimate the causal effect of aneurysm-repair surgery on (i) overall and (ii) aneurysm-related mortality in patients from the Effective Treatments for Thoracic Aortic Aneurysms (ETTAA) cohort. CCW artificially creates patient 'clones' labelled as surgery or no surgery within a 12-month grace period, and adjusts for informative censoring, thereby mitigating biases.

Results

Estimated seven-year survival if all patients have surgery was 57.4% (95% CI: 47.3%, 67.4%) compared to 49.9% (44.0%, 55.0%) if no patients had surgery. The benefit was primarily due to reduced aneurysm-related deaths (risk difference -8.7%, 95% CI -14.0%, -3.9%) with no significant effect on other causes.

Discussion

This study demonstrates trial emulation's feasibility for causal inference in the presence of competing risks. Findings suggest survival benefit from aneurysm-repair surgery to seven years post-enrolment to ETTAA. This approach targets causal effects of surgical interventions when RCTs are infeasible.

Maringe C, Majano SB, Exarchakou A, et al. (2020). Reflection on modern methods: trial emulation in the presence of immortal-time bias. Assessing the benefit of major surgery for elderly lung cancer patients using observational data. International Journal of Epidemiology, 49, 1719-29.

Murray J, Chesang C, Large S, et al. (2025). Trial emulation to assess the effect of surgery on survival when there are competing risks, with application to patients with thoracic aortic aneurysms. Journal of Clinical Epidemiology, 181.

Treatment trajectory extraction and balancing for longitudinal confounding control in target trial emulation from the French national health data system, a tocilizumab case-study.

14:30-14:45

Presenter: Hugo Malafosse

Authors: Hugo Malafosse [1]; Sarah Zohar [2]; David Pau [3]; Claire Castagné [3]; Jean Feydy [4]; Anne-Sophie Jannot [4]

Affiliations: [1] INSERM/INRIA HEKA team, Paris; [2] Inserm, Université Paris Cité, Inria, UMR 1346, HeKA, Paris; [3] Roche SAS, Boulogne-Billancourt, France, Boulogne-Billancourt; [4] HeKa Université Paris Cité, INRIA, INSERM, F-75015, Paris

Introduction

When starting a new treatment, patients accumulate complex treatment histories that proxy unmeasured confounders such as disease severity or treatment response. Standard target trial emulation approaches either discard this history (baseline propensity score PS) or compress it into scalar summaries (high-dimensional PS), destroying temporal information. We propose defining pairwise distance metrics directly on treatment sequences, treating them as structured objects rather than covariate vectors. We apply this to three tocilizumab trials using the French National Health Data System (SNDS).

Methods

From SNDS, daily drug dispensation time series are segmented into ordered therapeutic regimen periods via a change-point detection algorithm applied to multivariate dose exposure curves. Each segment is classified into a named therapeutic regimen using combined statistical and clinical rule-based methods. Three complementary distance metrics capture distinct dimensions of pre-t0 treatment similarity: Levenshtein on regimen label sequences (ordering), Needleman-Wunsch (temporal alignment with duration), and Euclidean distance on dose exposure distribution. These are combined via Mahalanobis distance and used for direct matching under the ATT estimand. We benchmark this sequence matching (C) against baseline PS (A) and L1-penalised high-dimensional PS incorporating >500 automated trajectory features (B). Proxy outcomes are mean glucocorticoid dose change for GiACTA (GCA) and OPTION (RA).

Results

Strategy B achieves the lowest overall standardised mean differences on scalar covariates, but within-pair distance metrics between patient's treatment history remain high, indicating poor trajectory similarity between compared patients. Strategy C produces 2-5× lower sequence distances and maintains stable covariate balance across pre-t0 history. The mean Levenshtein distance between patients for strategy B is 0.7, and 0.3 for C. This translates into significant differences in the treatment effect estimates between the methods: for corticosteroid dose change, B paradoxically reverses the expected signal (+2.1% GiACTA, +4.8% OPTION), while C detects significant reductions (-18.1% and -10.2%).

Discussion

Sequence distances preserve temporal structure that scalar summaries discard. Validation across three trials, two diseases and distinct endpoint types supports generalisability to sequential treatment settings in claims databases.

INVENTS has received funding from the European Union's Horizon Europe Research and Innovation program under grant agreement 101136365, SERI and the UKRI Innovative UK under their Horizon Europe Guarantee scheme.

Target trial emulation to evaluate the effect of insulin therapy in people with Cystic Fibrosis-Related Diabetes using European Cystic Fibrosis Society Patient Registry Data.

14:45-15:00

Presenter: Arianna Adamoli

Authors: Arianna Adamoli [1]; Annalisa Orenti [1, 2]; Simone Gambazza [1]; Federico Ambrogi [1, 3]; Anna Zolin [1]; Ruth Keogh [4]

Affiliations: [1] Università degli Studi di Milano, Milan; [2] IRCCS Ospedale Galeazzi-Sant'Ambrogio, Milan; [3] IRCCS Policlinico San Donato, San Donato Milanese; [4] London School of Hygiene & Tropical Medicine, London

Introduction

Cystic Fibrosis-Related Diabetes (CFRD) is the most common comorbidity in people with cystic fibrosis (pwCF). However, the association between insulin therapy (IT) for CFRD and lung function is unclear: observational studies suggest clinical benefits but are limited by small samples and confounding. This study aims to evaluate the effects of IT in pwCF with CFRD using data from the European Cystic Fibrosis Society Patient Registry (ECFSPR), including addressing the challenge of missing data on time-dependent confounders.

Methods

Using a Target Trial Emulation approach, we estimated the effects of sustained use of IT versus use of other CFRD therapies in pwCF aged ≥12 newly diagnosed with CFRD between 2018 and 2024, over 5 years of follow-up. The endpoint was (change in) percent predicted FEV_1 ($ppFEV_1$). To address missingness in time-dependent confounders, multiple imputation under a missing-at-random assumption was performed, taking into account longitudinal relationships and using a bootstrap resampling procedure. Specifically, bootstrap resampling was first applied, and multiple imputed datasets were generated for each resample. Marginal Structural Models with inverse probability weighting were then fitted on each imputed dataset to account for measured confounding and censoring, with estimates first pooled across imputations and then across bootstrap samples. Treatment effect modification by CFTR modulator use was assessed by including a three-way interaction term between time, IT use and modulator use.

Results

The study included 1,484 pwCF, of whom 61.3% initiated IT within the first year after CFRD diagnosis. Among individuals not using CFTR modulators, sustained IT use was associated with lower values of $ppFEV_1$ compared with alternative treatments (year 5: -2.38 95% CI: -5.76 to 0.99). Among CFTR modulator users, IT use was associated with better $ppFEV_1$ over shorter time periods (year 2: +1.41 95% CI: -0.18 to 3.01), but this difference was not sustained at year 5 (-1.29 95% CI: -4.24 to 1.66). All effect estimates were non-significant.

Discussion

In this registry-based study, IT was not associated with improved lung function compared with other CFRD treatments. Effect estimates showed modest differences by CFTR modulator use, with small, non-significant variations over time, suggesting limited evidence of heterogeneity in treatment response.

Research-Ready Data Integration and Software Workflows

Tue, 29.09.26, 13:30–15:00, Room: N2

The Biomedical Data Infrastructure - a National Database of Neurophysiological Data in Denmark

13:30-13:45

Presenter: Jonas Hügel

Authors: Jonas Hügel [1, 2]; Umaer Rashid Hanif [1, 2]; Sine Kongsbak Arvedsen [2]; Miki Nikolic [2]; Maria Joy Normann Haverberg [3]; Martin Ballegaard [3]; Tobias Søren Andersen [4]; Jakob Eyvind Bardram [1]; Poul Jørgen Jennum [2, 1]

Affiliations: [1] Department of Health Technology, Technical University of Denmark, Kongens Lyngby; [2] Danish Center for Sleep Medicine, Rigshospitalet, Glostrup; [3] Department of Neurophysiology, Zealand University Hospital, Roskilde; [4] Department of Applied Mathematics and Computer Sciences, Technical University of Denmark, Kongens Lyngby

Introduction

Neurophysiological recordings such as electroencephalography (EEG) and polysomnography (PSG) provide rich information on brain function and sleep, but their use in large-scale research is limited by heterogeneous data formats, fragmented storage, and restricted accessibility. The Biomedical Data Infrastructure (BDI) project aims to enable large-scale, harmonized use of neurophysiological data. By integrating clinical recordings across regions and linking them to national health registries, BDI hereby seeks to support research into sleep and neurological disorders and facilitate biomarker discovery.

Methods

The BDI project is divided into two phases. In the first phase, neurophysiological recordings and corresponding clinical reports from 2005 to 2023 are collected across the five Danish health management regions. Data is exported from clinical systems and transferred into an Secured Processing Environment (SPE), where harmonization is performed, including conversion of recordings into the open European Data Format (EDF+) standard. Additionally, we are exploring the use of locally running large language models (LLMs) to extract structured information from the corresponding reports.

In the second phase, we aim to provide researchers access to the harmonized data within the SPE. Based on a requirement elicitation process, we identified relevant use cases and requirements to select an appropriate host for the SPE.

Results

Legal agreements between the regions are being established to enable data collection. To date, ~120,000 recordings and reports have been extracted from two regions, consisting of ~140 TB of data. More than 20,000 recordings have been converted to EDF+ in collaboration with the Norwegian VIKING project.

15 key use cases have been identified, which served as the defining system requirements. The most critical requirements relate to access management (requesting/granting/removing), data analysis (graphical viewers for individual recordings and computational resources for training Machine Learning models), and integration with other registries, especially the statistical services in Denmark.

Discussion

BDI establishes a scalable, secure infrastructure that enables nationwide, registry-linked neurophysiological research. By transforming heterogeneous clinical data into harmonized, research-ready formats, the project provides a foundation for large-scale studies of sleep and neurological disorders and supports the identification of clinically relevant biomarkers. International collaboration is essential for addressing remaining technical challenges, particularly in data standardisation.

From Routine Healthcare Data to Research-Ready Cohorts: A Structured Preprocessing Framework

13:45-14:00

Presenter: Matthias Katzensteiner

Authors: Matthias Katzensteiner; Darian Liehr; Oliver J. Bott

Affiliations: Hochschule Hannover, Hannover

Introduction

The secondary use of routine clinical data is gaining importance due to national and international initiatives promoting data-driven research. In contrast to prospectively collected study data with predefined protocols, routine data are heterogeneous, irregular, and lack standardization, posing challenges for secondary use. Recent developments such as health data spaces and real-world evidence (RWE) highlight the need for structured preprocessing approaches to enable data reuse [1].

Methods

This work presents a structured preprocessing framework developed in the context of kidney transplant rejection research. The approach transforms routine healthcare data into research-ready datasets using a three-step algorithm: (A1) recursive event classification (rEC), (A2) context-sensitive event annotation (cEA), and (A3) context-based time series harmonisation (cTSH) [2]. The framework enables identification of clinically relevant events, their annotation to patient data, and transformation of irregular measurements into a standardized temporal grid.

Results

The framework enabled the construction of patient timelines and comparable datasets. In a proof-of-concept study (n = 141), three cTSH-based datasets were compared to a non-harmonised baseline using a random forest classifier. Predictive performance improved: accuracy increased from 82.9% to 88.6%, specificity from 64.3% to 72.7%, and precision from 80.0% to 88.5%, while sensitivity remained high (91.7-95.8% vs. 95.2%) [2]. In contrast, performance decreased when applying a Naive Bayes classifier, indicating model-dependent effects.

Discussion

The results demonstrate that the framework enables transformation of heterogeneous routine data into comparable datasets resembling clinical study data. The cTSH approach can be interpreted as a bridging technology, facilitating the use of interpretable machine learning models on complex routine data, particularly relevant in regulated contexts such as the Medical Device Regulation (MDR). Future work should compare this approach with statistical methods and deep learning models. The framework is transferable and contributes to secondary data research and RWE generation.

References

[1] Kympouropoulos S. Real World Evidence: methodological issues and opportunities from the European Health Data Space. BMC Med Res Methodol 202323(1):185. https://doi.org/10.1186/s12874-023-02014-3.

[2] Liehr D, Katzensteiner M, Bott OJ. On the Harmonisation of Time Series Data for the Optimisation of Machine Learning Using the Example of Rejection Prediction After Kidney Transplantation. Stud Health Technol Inform 2025327:68-72. https://doi.org/10.3233/SHTI250275.

The full paper of this abstract will be published in the GMDS series "Studies in Health Technology and Informatics" with doi: 10.3233/SHTI260978, URL:https://dx.doi.org/10.3233/SHTI260978

Detecting careless survey responses: automated data quality workflows using dataquieR

14:00-14:15

Presenter: Elena Salogni

Authors: Elena Salogni; Elisa Kasbohm; Stephan Struckmann; Carsten Oliver Schmidt

Affiliations: Universitätsmedizin Greifswald, Greifswald

Introduction

Careless or otherwise low-effort responding can distort survey-based epidemiologic evidence. Direct screening approaches such as bogus or instructed-response items are not always available, feasible, or desirable in established questionnaires. Indirect post hoc indices therefore offer a practical alternative because they are unobtrusive and can be applied to existing survey data. Empirical evidence suggests that several indirect indices perform at least as well as, and in some settings better

than, direct measures for detecting careless responding [1]. However, there is a lack of tools that integrate such measures into a broader data quality assessment workflow. For this reason, the R package dataquieR was extended by indirect indices to explicitly assess survey response quality.

Methods

We extended the R package dataquieR, to support automated assessment of survey response quality. The implementation covers indirect indicators from three domains: response invariability (maximum longstring, intra-individual response variability), response time (total completion time, time per item, relative completion speed), and response inconsistency (Mahalanobis distance). User-supplied metadata specify which item groups are eligible for which checks and define admissible ranges and grading rules. These specifications are provided in a spreadsheet and are used to generate standardized quality assessments within the dataquieR reporting workflow.

Results

The new functionality produces HTML-based output integrated into the dataquieR report under a dedicated Scales section. For each prespecified item group, the workflow summarizes calculated response-quality indicators, flags observations outside expected ranges, and applies customizable grading rules. This enables transparent, reproducible, and study-specific screening of potential low-quality survey responses without modifying the questionnaire itself. Assessment results based on questionnaire data from the Study of Health in Pomerania (SHIP) are used to illustrate the application.

Discussion

The extension enables reproducible post hoc screening for problematic survey response patterns in epidemiologic studies and may help identify data requiring further review. However, utility in practice depends on appropriate metadata, instrument structure, and empirically justified decision thresholds.

[1] Goldammer, P., Stöckli, P.L., Escher, Y.A. et al. Careless responding detection revisited: Accuracy of direct and indirect measures. Behav Res 56, 8422-8449 (2024). https://doi.org/10.3758/s13428-024-02484-3

The DZHK Feasibility Explorer: Supporting Cohort Discovery and Feasibility Assessment for Cardiovascular Research

14:15-14:30

Presenter: Sabine Hanß

Authors: Eva-Maria Rieß [1]; Robert Kossen [1]; Dagmar Krefting [1, 2, 3]; Kevin Alejandro Molina [1]; Heiko Scheel [4, 2]; Sabine Hanß [5, 2]

Affiliations: [1] Department of Medical Informatics, University Medical Center Göttingen (UMG), Göttingen; [2] DZHK (German Centre for Cardiovascular Research), Partner Site Göttingen, Germany, Göttingen; [3] Campus Institute Data Science, Georg-August-University Göttingen, Göttingen; [4] Department of Medical Informatics, University Medical Center Göttingen (UMG), Götttingen; [5] University Medical Center Goettingen, Göttingen

Introduction

he secondary use of clinical study data and biological samples holds substantial potential for advancing biomedical research. The German Centre for Cardiovascular Research (DZHK) research platform made the secondary use of study data possible through standardized data collections and centralized data management. However, identifying suitable datasets and estimating cohort sizes prior to formal data access requests remains challenging due to restricted access to data and metadata. To address this gap, we developed the Feasibility Explorer (FE), a privacy-preserving tool that enables researchers interested in using the DZHK research data to explore data availability and assess study feasibility.

Methods

The FE was designed based on requirements for multi-study research environments, analysis of existing cohort discovery tools, and DZHK-specific governance and data protection policies. It follows a layered architecture comprising presentation, application, and data-processing layers, enabling distributed querying across heterogeneous data sources, including clinical data, imaging data, and biological samples. A web-based user interface allows intuitive query construction via filters and presents aggregated results using predefined visualizations.

Results

To ensure compliance with data protection regulations, including the General Data Protection Regulation (GDPR), the FE restricts outputs to aggregated data, suppresses small cohort sizes, and implements a blurring mechanism to mitigate inference attacks. Data integration is performed on demand using pseudonym mapping, and no individual-level data are exposed or persistently stored.

Discussion

Since its productive deployment, the FE has supported researchers in assessing data availability and estimating cohort sizes prior to submitting formal applications. While initial experiences indicate improved transparency and usability, formal evaluation is ongoing. The Feasibility Explorer provides a practical approach to support feasibility assessments in a federated research data environment, enabling efficient and informed study planning while maintaining strict data protection constraints.

The full paper of this abstract will be published in the GMDS series "Studies in Health Technology and Informatics" with doi: 10.3233/SHTI260985, URL:https://dx.doi.org/10.3233/SHTI260985

The MeDIC dashboard: a tool to visually communicate patient health care data

14:30-14:45

Presenter: Sherry Freiesleben

Authors: Sherry Freiesleben [1]; Carl Philip Röseberg [1]; Joe Kranzusch [2]; Kai Fitzer [2]; Dagmar Waltemath [2, 3]; Petra Gröber [1]

Affiliations: [1] Medical Data Integration Center, Rostock University Medical Center, Rostock; [2] Core Unit Datenintegrationszentrum, Universitätsmedizin Greifswald, Greifswald; [3] Department of Medical Informatics, Greifswald University Medical Hospital, Greifswald

Introduction

The Medical Data Integration Centers (MeDICs) Rostock and Greifswald are tasked with extracting and transforming clinical routine data and have already generated millions of FHIR resources for secondary use. However, an appropriate tool to properly communicate this information, and to compile, summarize, and visualize key performance data was lacking. Therefore, we developed a dashboard that displays key information, supports data exploration and facilitates first point of contact with MeDICs. The dashboard can be operated independently at several MeDIC locations, as we have shown for two data integration centers.

Methods

To this end, we used the EPIS framework to design an effective, sustainable, and impactful dashboard. We first determined who our end users were and then identified metrics of value to relevant actors. We accordingly designed our dashboard around the Medical Informatics Initiative core dataset and implemented it using R Shiny, thereby retrieving available FHIR resources using fhircrackr. Using an iterative feedback approach, we adapted the dashboard to better accommodate the needs of our users.

Results

Our web-based dashboard was implemented in German, tested in Rostock and then deployed at the MeDICs in Rostock and Greifswald. It automatically updates every 24 hours and allows for remote monitoring. The available FHIR data is organized and displayed using value boxes, visualizations such as bar and line charts, and searchable tables. Due to the varying levels of expertise and expectations among different end users, we generated a combination of an administrative and a clinical dashboard. Our dashboard was also valuable when communicating with other MeDICs we were easily able to show what type and amount of FHIR resources were already assessed and to what extent. Moreover, the dashboard is easily customizable and content can be added or changed as required by users.

Discussion

Overall, this project demonstrated that it is effective and convenient to communicate up-to-date information related to routine healthcare via a dashboard. It also showed that it is advantageous to have an easily customizable dashboard to quickly react to requests from end users. Furthermore, the use of harmonized and standardized data is paramount when creating a tool that can potentially be used at a national-level.

---

mmid: Multi-Modal Integration and Downstream analyses for healthcare analytics in Python

14:45-15:00

Presenter: Andrea Mario Vergani

Authors: Andrea Mario Vergani [1, 2]; Valeria Iapaolo [2]; Emanuele Di Angelantonio [1]; Marco Masseroli [2]; Francesca Ieva [2, 1]

Affiliations: [1] Human Technopole, Milano; [2] Politecnico di Milano, Milano

Introduction

mmid (Multi-Modal Integration and Downstream analyses for healthcare analytics) is a Python package that offers multi-modal fusion and imputation, classification, time-to-event prediction and clustering functionalities under a single interface, filling the gap of sequential data integration and downstream analyses for healthcare applications in a structured and flexible environment.

Methods

mmid wraps in a unique package several algorithms for multi-modal decomposition (such as MOFA+ and AJIVE), prediction (i.e., disease classification and time-to-event analysis) and clustering (e.g., DBSCAN), which can be combined smoothly with a single command and proper configuration files, thus facilitating reproducibility and transferability of studies involving heterogeneous health data sources. A showcase on personalised cardiovascular risk prediction is used to highlight the relevance of a composite pipeline enabling proper treatment and analysis of complex multi-modal data. We thus employed mmid in a real application scenario involving cardiac magnetic resonance imaging, electrocardiogram, and polygenic risk scores data from the UK Biobank.

Results

Thanks to mmid, we smoothly proved that the three modalities captured joint and individual information that was used to (1) early identify cardiovascular disease before clinical manifestations with cardiological relevance, and (2) do it better than single data sources alone. Moreover, mmid allowed to impute partially observable data modalities without considerable performance losses in downstream disease prediction, thus proving its relevance for real-world health analytics applications (which are often characterised by the presence of missing data). All the analyses and comparisons between model and data runs of our application study were easily performed with the mmid Python package, just properly modifying input parameters and configuration files.

Discussion

mmid bridges the gap between standalone multi-modal integration algorithms (and corresponding libraries) and the need to compare their outputs and use them quickly and easily for a variety of disease prediction and/or patient clustering analyses. The package enables access to state-of-the-art fusion methods in a unified way, guaranteeing transferability, easier reproducibility, and efficiency of two-step (i.e., integration plus downstream) multi-modal analyses. Finally, mmid is extensible,

allowing novel data fusion and downstream algorithms to be easily included in its future versions. The package proved its relevance in a real application study involving high-dimensional health data.

Prediction and Treatment

Tue, 29.09.26, 13:30–15:00, Room: N3

Chair: David van Klaveren

Debiased machine learning for combining probability and non-probability survey data

13:30-13:45

Presenter/Author: Shaun Seaman

Affiliations: University of Cambridge, Cambridge

Introduction

Nonprobability samples, e.g. web-based volunteer surveys, have the advantage of often being readily available or relatively cheap to obtain. However, such samples may not be representive of the population. Methods have been proposed for using data from a probability survey to address non-representative of a nonprobability sample (e.g Chen et al., 2020). We shall consider the problem of estimating the finite population mean Ybar of an outcome Y using data on Y from a nonprobability sample and auxiliary information from a probability sample. Existing double robust (DR) estimators of Ybar require the estimation of two nuisance functions: the conditional probability of selection into the nonprobability sample given auxiliary variables X that are observed in both samples, and the conditional expectation of Y given X. These nuisance functions can be estimated using parametric models, but the resulting estimator of Ybar will typically be biased if both parametric models are misspecified, as would often be the case. It would therefore be advantageous to be able to use more flexible data-adaptive / machine-learning estimators of the nuisance functions.

Methods

We develop a general framework for the valid use of DR estimators of Ybar when the design of the probability sample uses sampling without replacement at the first stage and data-adaptive / machine-learning estimators are used for the nuisance functions. We prove that several DR estimators of Ybar, including targeted maximum likelihood estimators, are asymptotically normally distributed with an easily estimated variance when the estimators of the nuisance functions converge faster than the n^{1/4} rate and cross-fitting is used.

Results

We present a simulation study that demonstrates good performance of these DR estimators compared to the corresponding estimators that rely on at least one correctly specified parametric model. Our methods will be applied to data from the United Kingdom's third National Survey of Sexual Attitudes and Lifestyles.

Discussion

The proposed methods reduce the risk of biased estimation of a population mean caused by model misspecification.

References:

1. Seaman SR. Debiased machine learning for combining probability and non-probability survey data. arXiv:2508.08948.

2. Chen, Li, and Wu. Doubly robust inference with nonprobability survey samples. JASA, 115:2011-2021, 2020.

Constrained maximum log-likelihood estimation for adjusted Risk Differences (aRD)

13:45-14:00

Presenter: Stefan Wagenpfeil

Authors: Stefan Wagenpfeil [1]; Julius Johannes Weise [2]; Thomas Wolf [3]

Affiliations: [1] Universität des Saarlandes, Institut für Medizinische Biometrie, Epidemiologie und Medizinische Informatik, Homburg (Saar); [2] Institut für Medizinische Biometrie, Epidemiologie und Medizinische Informatik, Universität des Saarlandes, Homburg, Homburg; [3] Universität des Saarlandes, Institut für Medizinische Biometrie, Epidemiologie und Medizinische Informatik, Homburg

Introduction

Adjusted risk differences (aRD) can be modelled using linear-binomial regression (LiB) for prospective epidemiological studies. The difference to logistic regression (LoR) is only the response function, i.e. the link between the linear predictor (LP) and the probability of one out of two outcome categories. While the response function for LoR is the logistic distribution, it is simply the identity (ID) in LiB. However, with unconstrained maximum log-likelihood estimation (ML), this can lead to an even complex-valued objective function due to the natural logarithm of a possibly negative or LP >1, which is unfeasible.

Methods

For the constrained maximum likelihood (cML) approach in LiB modelling we developed a) adequate linear box constraints defining the underlying nonlinear programming problem for the purpose of prediction, and b) constructed a valid starting vector of covariate effects and baseline parameter for cML.

Results

We present a cML approach that can be seen as a modification of Newton's method allowing for linear inequality box constraints. Considering constraints, we focus on the construction of box constraints for the purpose of prediction. More details on numerical properties of global convergence

and locally quadratic convergence speed can be found in [1]. A real data example as well as results from simulated data are presented. The algorithm is implemented as aRD in the R CRAN package.

Discussion

A similar problem with possibly unfeasible ML was presented in [2] for modelling adjusted relative risks. This issue was fixed in [3] and presented on the IBC 2024 in Atlanta, USA. Now, the adequate statistical modelling approach for adjusted relative risks as well as risk differences can be used even for large-scale prospective, epidemiological studies without looking for numerical workarounds.

[1] Kredler Ch (1996) A SQP-Method for Linearly Constrained Maximum Likelihood Problems, in App. Math. and Paral. Comp. - Festschr. for K. RITTER, pp. 157-174. Physica, Springer, Heidelberg.

[2] Williamson T., et al.: Log-binomial models: exploring failed convergence, Emerg. Themes Epidemiol. 2013(10:14),1-10.

[3] Wagenpfeil S et al. (2024) Estimation of adjusted relative risks in log-binomial regression using the BSW algorithm. Biokybernetika, Ch. 31, de Gruyter, Boston.

Predicting heterogeneous treatment effects from randomised trials: a comparison of the risk modeling and effect modification approaches

14:00-14:15

Presenter: Frederik Philipona

Authors: Frederik Philipona [1]; Georgia Salanti [1]; Orestis Efthimiou [1]; Marta Mainetti [2]

Affiliations: [1] Institute of Social and Preventive Medicine, University of Bern, Bern; [2] University of Bern, Bern

Introduction

Risk modeling (RM) and treatment effect modification (EM) are two approaches to build models to predict heterogeneous treatment effects using data from randomized control trials (RCT). RM is a two-stage approach that predicts baseline risk at the first stage and then employs it as the only effect modifier in a model fit at the second stage. Conversely, with EM we build multivariable models including all relevant patient characteristics. A thorough comparison between the approaches is missing. We hereby present the assumptions underlying each approach, discuss potential limitations, and conduct an extensive simulation study to compare them.

Methods

We unravel the assumptions underpinning the approaches, discuss the likelihood that these assumptions are fulfilled, and present their theoretical relative advantages. In simulations, we explore scenarios for a dichotomous outcome where the RM and EM approaches are implemented with penalized logistic regression models. We evaluate the performance of both approaches with respect to the mean square error in the predicted risk difference.

Results

In the theoretical part, we show that baseline risk is a treatment effect modifier in many clinical scenarios. RM is a dimensionality reduction approach which is based on that fact. However, it makes strong assumptions about the role of prognostic factors modifying the treatment effect. EM's greater flexibility is a possible advantage when the sample size is large.

In most simulated scenarios, RM performed better than EM, even in scenarios where the assumptions underlying RM were not met. The advantage of RM diminished as sample size increased. Further, when the outcome was rare and the sample size was small, EM outperformed RM.

Discussion

This work compares theoretically and in simulations RM and EM for the prediction of HTEs from RCT data. We show that both approaches have their advantages and disadvantages. Based on our simulations RM outperforms EM for sample sizes small to moderate (<2000 participants), which are common in research. For larger sample sizes the difference between the two approaches is small.

Free-of-Imputation Regression Forest to Handle Missingness in Predicting Individual Treatment Effects

14:15-14:30

Presenter: Pamela Solano

Authors: Pamela Solano [1]; Thomas Jaki [2]

Affiliations: [1] Faculty of Computer Science and Data Science, University of Regensburg, Germany, Regensburg; [2] Faculty of Computer Science and Data Science, University of Regensburg, Germany. MRC Biostatistics Unit, University of Cambridge, UK, Regensburg

Introduction

Missing data are common in applied research and remain a persistent challenge in statistical methodology. In precision medicine, failure to handle missingness appropriately may compromise the validity of conclusions. In particular, clinicians relying on evidence from randomized clinical trials may encounter difficulty when translating population-level results into treatment decisions for individual patients, even within a coherent and reliable framework.

Standard approaches to missing data typically rely on multiple imputation, followed by combining multiple estimates across completed datasets. However, missing data can introduce bias and distort statistical inference, and even well-established methods do not ensure unbiased inference when the missingness mechanism departs from the Missing Completely at Random (MCAR) assumption. These concerns are amplified in the context of predicting individual treatment effect (PITE), defined as the difference between predicted outcomes under treatment and control. Estimation of PITE is inherently challenging due to unobserved counterfactuals, high-dimensional covariate spaces, and

complex interactions. Furthermore, the extent to which different strategies for handling missingness affect predictive performance - in terms of bias, variability, and computational cost - remains insufficiently understood.

Methods

In this work, we treat missingness as potentially informative rather than as a nuisance to be removed. We propose a regression forest modeling approach that incorporates missingness directly, thereby avoiding reliance on multiple imputation. The aim is to improve statistical efficiency and reduce bias in predictive performance.

We conduct simulation studies based on two-arm randomized clinical trials with continuous outcomes. The design considers Missing at Random (MAR) and Missing Not at Random (MNAR) mechanisms in covariates, along with high-dimensional settings, complex interactions, and varying proportions of missingness.

Results

Our results indicate that the choice of imputation method, model specification, and missingness handling strategy materially affect both the accuracy and the uncertainty of PITE estimates by inducing propagation of bias and inflating the variance of estimated treatment effects. In contrast, regression tree methods that do not rely on imputation achieve improved computational efficiency, lower prediction error (under MCAR), and perform comparably to standard multiple imputation approaches under MAR and MNAR.

Discussion

Tree-based models that incorporate missingness directly provide a practical alternative to multiple imputation for estimating individual treatment effects with missing covariates.

A simulation study investigating the impact of clinical prediction model-based treatment initiation on model updating

14:30-14:45

Presenter: Samantha Pacynko

Authors: Samantha Pacynko; Matthew Sperrin; David Jenkins

Affiliations: Division of Informatics, Imaging and Data Sciences, School of Health Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester, UK, Manchester

Introduction

When clinical prediction models (CPMs) are updated, there is a potential feedback loop if data for individuals whose treatment was influenced by the model are included in the updating process. Variables previously predictive of a negative outcome become associated with treatment and improved

outcomes: the 'prediction paradox'. This can cause reduced model performance and suboptimal decision support over time, but there is limited evidence on the magnitude of this deterioration. The aim of this work is to quantify the effect of the prediction paradox across a range of different scenarios.

Methods

A simulation study was conducted in which CPMs, developed to predict the probability of a binary adverse outcome, were used to guide treatment allocation and then subsequently updated. Parameters were varied to create different scenarios: 1) whether treatment was included in the CPM, 2) presence of an unmeasured confounder, 3) outcome prevalence, 4) treatment allocation threshold, 5) risk reduction with treatment. Several measures of predictive performance were calculated: calibration-in-the-large (CITL), calibration slope, c-statistic, Brier score, and count of incorrect treatment decisions.

Results

After being updated, CPMs which did not include treatment as a predictor variable underestimated risk and showed poor predictive accuracy when used to make predictions for untreated individuals across all scenarios. The worst performance led to a final average CITL of 1.19 and only 75% of individuals receiving predictions that would correctly identify if they should receive treatment or not. Other scenarios were less extreme but still not ideal generally performance was worse when risk reduction of treatment was higher or when treatment allocation threshold was lower than outcome prevalence. Including treatment as a predictor improved performance, but risk was still underestimated when an unmeasured confounding variable was present.

Discussion

When a model guides treatment allocation, model updating can cause underestimation of risk, missed opportunities for treatment, and adverse outcomes that could have been avoided. Including treatment as a predictor can reduce the magnitude of this effect but is unlikely to fully mitigate it in real-world data due to the likelihood of unmeasured confounding variables. More advanced modelling approaches need to be investigated in relation to this problem area.

USING PKPD MODELLING AND EXTERNAL EVIDENCE TO INFORM PAEDIATRIC VIRTUAL PATIENTS GENERATION: A CASE STUDY

14:45-15:00

Presenter: Tarini Singh

Authors: Tarini Singh [1]; Meemansa Sood [2]; Letao Li [1]; Sarah Zohar [1]; Moreno Ursino [1]; Marina Savelieva [2]

Affiliations: [1] Inserm, Université Paris Cité, Inria, UMR 1346, HeKA, Paris; [2] Novartis Pharma AG, Basel

Introduction

The INVENTS project aims to strengthen modelling and simulation approaches and their regulatory acceptance to accelerate drug development in rare diseases. Paediatric multiple sclerosis (MS) is rare, representing only 3-5% of MS cases worldwide, predominantly of the relapsing-remitting MS (RRMS) subtype. This work evaluates whether population PK and PKPD models developed from anonymised adult data and external evidence can be qualified to generate virtual paediatric RRMS patients (vpRRMSp).

Methods

Novartis Pharma AG granted access to anonymised phase III fingolimod trials data from 1856 adults and 106 paediatric patients treated once-daily for two years. The modelling strategy followed ICH M15 and credibilityframework from Musuamba et al. Post-anonymisation covariates included sex, age, weight, race (Caucasians/ Asians/ Others), MS duration (3-year intervals) and prior treatment. Impact of anonymisation was assessed bycomparing results with PK analyses pre-anonymisation. A two-compartment NLME model with time-lag, first-order absorption, clearance and volumes of distribution (with Phase I Healthy Volunteers) was evaluated in adults. Model-predicted concentrations were used to inform a direct inhibitory Emax model for Absolute Lymphocyte Counts (ALC), a key RRMS biomarker. Covariate and error model selection used the Stochastic Approximation Model Building Algorithm, and model adequacy was assessed using Goodness-of-fit, RSEs, and shrinkage. To evaluate suitability for vpRRMSp generation, adult PKPD parameters were used to simulate concentrations and ALC responses for 1000 paediatric individuals. The model was then refitted to observed paediatric data using penalised maximum likelihood estimation (PMLE) and pooled analysis, and validated with 5-fold cross-validation.

Results

Predictions from anonymised data were consistent with pre-anonymisation PK analyses. Model built using Phase I adult Healthy Volunteers performed well in Phase III MS patients (RSEs <30% for model parameters and covariates). Simulation-based verification using the adult PK and PKPD model showed most paediatric observations within the 5-95% prediction intervals. Additionally, the model performed well when refitted to paediatric data and passed the validation checks for PMLE and pooled analysis.

Discussion

These results represent initial steps in establishing model credibility for generating vpRRMSp. Future work will refine extrapolation and extend analyses to additional efficacy endpoints to support virtual control-arm development aligned with evolving regulatory expectations.

Treatment Effect Heterogeneity and Clinical Interpretation

Tue, 29.09.26, 13:30–15:00, Room: N5

Chair: Katja Ickstadt

Modelling varying treatment effects using proportional interactions

13:30-13:45

Presenter: Leon Di Stefano

Authors: Leon Di Stefano [1, 2]; Ravi Varadhan [3]; David Wakeham [4]; Elizabeth Stuart [5]

Affiliations: [1] Murdoch Children's Research Institute, Melbourne; [2] University of Melbourne, Melbourne; [3] Johns Hopkins School of Medicine, Baltimore; [4] Torsor Labs, Vancouver; [5] Johns Hopkins University Bloomberg School of Public Health, Baltimore

Introduction

Estimating varying treatment effects in clinical trials is hard, and the standard approach - multiple subgroup analyses - leads to difficulties with both inference and interpretation. The proportional interaction model is a one-parameter extension of additive regression adjustment in which a single linear "score" predicts both average outcomes and varying treatment effects (Follmann and Proschan,1999 Kovalchik, Varadhan, and Weiss, 2013). A popular two-step approach estimates a prognostic score and then evaluates its interaction with treatment (Kent et al, Trials 2010).

Methods

We develop one-step, likelihood-based inference for the proportional interaction model. An angular reparameterization respects symmetries of the model and stabilizes fitting. Asymptotic analysis reveals how the accuracy with which proportional interactions can be estimated depends on the prognostic strength of the covariates. We conduct simulation studies assessing the performance of our approach both when proportional interactions are present and under misspecification. Finally, we re-analyse a clinical trial of schizophrenia treatments to show how our method can be incorporated into a traditional model-comparison framework.

Results

Proportional interaction coefficients of 2/3 or 1.5 are reliably detectable (80% power, 5% false positive rate) with sample sizes of approximately 500 to 2000 as R2 among controls varies from 0.3 to 0.1. Our approach performs well against two-step methods, especially when the prognostic score predicts better in the treated group. Performance breaks down, however, when R2 is small relative to sample size and when there are more (20) rather than fewer (5) covariates. Estimating proportional interactions leads to higher discrimination-for-benefit compared to keeping only significant interactions under moderate misspecification.

Discussion

The proportional interaction model can be stably fit and may be a useful complement to multiple subgroup analyses when prognostic information is reasonably strong.

References:

1. Kovalchik SA, Varadhan R, Weiss CO. Assessing heterogeneity of treatment effect in a clinical trial with the proportional interactions model. Statistics in Medicine. 201332(28):4906-4923. doi:10.1002/sim.5881

2. Follmann DA, Proschan MA. A Multivariate Test of Interaction for Use in Clinical Trials. Biometrics. 199955(4):1151-1155. doi:10.1111/j.0006-341X.1999.01151.x

3. Kent DM, Rothwell PM, Ioannidis JP, Altman DG, Hayward RA. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. Trials. 201011(1):85. doi:10.1186/1745-6215-11-85

A Bayesian perspective on predictive approaches to heterogeneity of treatment effects: a case study of the IST-3 trial

13:45-14:00

Presenter/Author: Markus Huber

Affiliations: Department of Anaesthesiology and Pain Medicine, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland, Bern

Introduction

The statistical analysis of randomized controlled trials provides a crucial step in informing clinicians, sponsors and regulatory agencies regarding the efficacy and safety of two or more alternative treatments. Over the past decades, various statistical methods were proposed and re-discovered to overcome the limitations of traditional trial analyses. These methods feature predictive approaches to heterogeneity of treatment effects and Bayesian reanalyses. Here, we combine these two approaches to examine the degree of evidence regarding the superiorty of Alteplase (rt-PA) compared to placebo in patients suffering acute stroke.

Methods

Data from the third international stroke trial (IST-3) was re-analyzed with Bayesian methods and risk-modelling approaches. For three standard priors, the probability of rt-PA superiority was examined both in the full trial population as well as in quartile-based risk-subgroups. In these risk-subgroups, patients were grouped according to their baseline risks of being alive and independent at 6-months follow-up (primary outcome). The baseline risks were computed with a multivariable logistic regression model fitted with trial data.

Results

The IST-3 trial provided evidence of overall rt-PA superiority with a probability of 73%-96%, in particular for patients with a low baseline risk of a favorable outcome at 6-months follow. In contrast, the probability for rt-PA superiority is low in patients that are likely to experience a favorable outcome (without rt-PA treatment). Trial data before 2007 at the time of sample size and effect size reconsiderations already showed evidence of heterogeneity of treatment effects (p=0.026). While there is evidence of an overall rt-PA superiority, we found only low probability of a treatment benefit with respect to effect sizes of 3%, 4.7% and 10% which were considered by the investigators.

Discussion

While the null hypothesis of the IST-3 trial could not be rejected from a frequentist perspective (p=0.181), this study demonstrated a nuanced probability of both benefits and harms of rt-PA compared to placebo. We conclude that a Bayesian perspective on predictive approaches to heterogeneity of treatments effects provides a clinically relevant approach to examine the degree how assumptions made during trial planning apply to groups of patients sharing similar risks of experiencing the primary outcome.

Methods for clinical risk prediction under interventions: a comparative simulation study

14:00-14:15

Presenter: Andriana Kostouraki

Authors: Andriana Kostouraki; Elizabeth Williamson

Affiliations: London School of Hygiene and Tropical Medicine, London

Introduction

When an effective intervention exists, individuals deemed high-risk are likely to receive it. Developing a risk prediction model on a data set of individuals intervention-free at baseline might be affected by later intervention 'drop-in', i.e. high-risk patients preferentially receiving the intervention. This may lead to underestimation of risk under no intervention for high-risk patients. Recent developments which blend causality concepts into risk prediction modelling (often called predictions under interventions), correctly allow for this setting, but require stronger assumptions and more complex modelling strategies. Additionally, it is unclear what the comparative performance of these developments is.

Methods

Approaches to account for intervention drop-in, could be grouped into those based on modelling i. the intervention, ii. the outcome mechanism - both conditional on intervention, covariates and outcome histories (propensity score and g-formula related, respectively), iii. both the intervention and outcome, iv. the outcome conditional on predictors and incorporating a time-dependent intervention effect, pooled from an external source, v. those not accounting for intervention drop-in,

referred to as 'standard'. We perform simulations to evaluate and compare the methods in longitudinal observational settings with time-varying confounding and intervention strategies sustained over time i.e. 'intervention always' vs 'no intervention always'. We focus on time-to-event outcomes. We explore scenarios where the development and validation data may/may not differ and apply the Brier score, the area under the curve and the calibration slope, adapted to assess predictive performance under 'intervention always' and under 'no intervention always', as recently proposed in the literature. To estimate the true predictive performance, we generate 'perfect' counterfactual validation data sets.

Results

Accounting for intervention drop-in enables better prediction under intervention performance the 'standard' approach performs worse when the target is prediction of risk under 'no intervention always'.

Discussion

When predicting risk under intervention given predictors, causal inference offers well-established methods to correctly target the population of interest. Evaluation measures must appropriately assess the performance in predicting risk under interventions therefore, standard measures must be adapted accordingly to contrast the predicted outcomes against the outcomes that would have been observed had individuals been assigned the intervention of interest.

Estimating the probability of experiencing an improvement in randomized controlled trials based on anchor-based MIC studies - an alternative to responder analyses

14:15-14:30

Presenter: Werner Vach

Authors: Werner Vach [1, 2]; Franziska Saxer [3, 4]

Affiliations: [1] Basel Academy for Quality and Research in Medicine, Basel; [2] Department of Sport Sciences and Clinical Biomechanics, University of Southern Denmark, Odense; [3] Novartis Biomedical Research, Basel; [4] Faculty of Medicine, University of Basel, Basel

Introduction

When using a patient reported outcome as primary endpoint in a randomized contolledl trial (RCT), responder analyses are widely used to describe the relevance of an observed effect attributable to the intervention of interest. The choice of thresholds in responder analyses is often based on minimal important change (MIC) values.

Methods

As an alternative to simply computing MIC values, we suggest determining a translation function allowing to translate change scores values into the probability of experiencing an improvement. The application of such a translation function to the change scores observed in an RCT allows estimating arm-specific probabilities of experiencing an improvement. These values are conceptually similar to responder frequencies. The approach is illustrated using a pair of synthetic studies, both constructed mimicking existing studies. A simulation study investigates the gain in efficiency from avoiding the dichotomization.

Results

The illustrative application of the methodology demonstrates that the approach is feasible and allows drawing conclusions in a similar way as in a responder analysis, but with greater precision. The simulation study confirms that this gain in efficiency holds more generally. Considering all response levels of the anchor variable may allow gaining more nuanced insights.

Discussion

Estimating the probability of experiencing an improvement provides an alternative to responder analyses. This approach makes more efficient use of information comprised in the data from anchor-based MIC studies than the translation into MIC values to inform the choice of thresholds in a responder analysis.

References:

Vach W, Saxer F. Estimating the probability of experiencing an improvement in randomized controlled trials based on anchor-based MIC studies - An alternative to responder analyses. Contemp Clin Trials. 2026 162:108243. doi: 10.1016/j.cct.2026.108243.

Unified implementation and comparison of Bayesian shrinkage methods for treatment effect estimation in subgroups

14:30-14:45

Presenter: Marcel Wolbers

Authors: Marcel Wolbers; Miriam Pedrera Gómez; Alex Ocampo; Isaac Gravestock

Affiliations: F. Hoffmann - La Roche Ltd, Basel

Introduction

Evaluating treatment effect heterogeneity across patient subgroups is a fundamental aspect of clinical trial analysis. However, these analyses face inherent limitations due to small sample sizes and the multiplicity of subgroups. There is also a tendency to overinterpret extreme estimates, which often reflect random variation rather than true effects, potentially causing spurious clinical conclusions. As an alternative to conventional methods, statisticians have explored Bayesian shrinkage methods,

which use priors to shrink raw treatment effect estimates in subgroups toward the overall effect. While various shrinkage estimators exist, it remains unclear which ones perform best.

Methods

This work provides a unified presentation, a software implementation (in the R package bonsaiforest2), and a simulation comparison of one-way and global shrinkage methods for continuous, binary, count, and time-to-event endpoints. One-way models fit separate shrinkage models for each subgrouping variable, whereas global models include all subgroup indicators simultaneously. Both approaches derive standardized, subgroup-specific treatment effects via G-computation.

Results

Across the simulation scenarios, shrinkage methods consistently outperformed standard estimators without shrinkage in terms of mean squared error. They were also more efficient at identifying non-efficacious subgroups. Global shrinkage models (using a regularized horseshoe prior) generally showed a smaller mean squared error and less dependence on hyperprior parameters than one-way models (using normal priors with a half-normal hyperprior for the standard deviation), but exhibited slightly larger bias and worse frequentist coverage of credible intervals. For both approaches, hyperprior choices anchored in trial assumptions about the anticipated overall treatment effect size performed well.

Discussion

Some degree of shrinkage is preferable to none. We advocate for routinely including shrunken estimates in clinical forest plots to facilitate robust decision-making. Our flexible R package bonsaiforest2 (https://openpharma.github.io/bonsaiforest2/) supports this practice.

Reference

Wolbers, Marcel, Miriam Pedrera Gómez, Alex Ocampo, and Isaac Gravestock. 2026. "Unified Implementation and Comparison of Bayesian Shrinkage Methods for Treatment Effect Estimation in Subgroups." arXiv Preprint arXiv:2603.21967. https://arxiv.org/abs/2603.21967.

Benefit-Risk Assessment of Non-Inferiority Trials Using Generalized Pairwise Comparison

14:45-15:00

Presenter: Youmei Han

Authors: Youmei Han [1]; Koji Oba [2, 3]; Tomohiro Shinozaki [3, 2]; Yutaka Matsuyama [3]

Affiliations: [1] School of Integrated Health Sciences, Faculty of Medicine, The University of Tokyo, Tokyo; [2] Interfaculty Initiative in Information Studies, The University of Tokyo, Tokyo; [3] Department of Biostatistics, School of Public Health, Graduate School of Medicine, The University of Tokyo, Tokyo

Introduction

Non-inferiority trials aim to prove the overall benefit of a new treatment. However, conventional non-inferiority trials analyze endpoints separately, hindering benefit-risk assessments. Generalized pairwise comparison (GPC) has emerged as a potential solution, but its operational characteristics for testing non-inferiority hypotheses remain unclear.

Methods

We conducted simulations using disease-free survival (DFS) as the primary and QOL as the secondary endpoint. Four base scenarios were considered: (A) DFS non-inferior and QOL superior (truly non-inferior) (B) DFS non-inferior without QOL superiority (C) DFS inferior with QOL superiority and (D) both inferior. Hazard ratio (HR), event rates, and QOL effect sizes were varied across 48 scenarios, each replicated 10,000 times. In each dataset, GPC constructs pairwise data between randomized groups, and firstly compares DFS using the "win/lose" threshold translated from the non-inferiority margin on the hazard ratio scale, followed by a QOL comparison. Non-inferiority was demonstrated if the lower limit of the 95% confidence interval for the net benefit (the difference between "win" and "lose" probabilities) was above 0, based on a Wald-type test. Performance metrics included the proportions of treatment adoption by GPC, HR for DFS from the Cox model, and hierarchical procedure test of DFS (Cox model) and QOL (t-test). GPC was also applied to the RESPECT trial of trastuzumab monotherapy in older patients with early breast cancer [1].

Results

In scenario (A), GPC maintained high adoption rates even with low event counts, where Cox regression and hierarchical procedure tests showed reduced power. In scenarios (B) and (D), GPC appropriately controlled type I error below 5%. However, in scenario (C), large QOL advantages could offset DFS inferiority, particularly when event rates were low. In the RESPECT trial, the net benefit favored trastuzumab monotherapy (0.20 95% CI, 0.05-0.34 P=0.008). Multiple sensitivity analyses confirmed the robustness of these results, suggesting the clinical benefit of monotherapy.

Discussion

GPC enables integrated benefit-risk assessment of non-inferiority trials and may overcome limitations of conventional HR-based approaches, especially in rare-event settings. Nonetheless, its capacity to offset survival disadvantages with secondary benefits necessitates careful clinical interpretation and pre-specification of acceptable trade-offs.

[1] Sawaki M, et al. J Clin Oncol. 202038(32):3743-3752.

Building and Updating Prediction Models

Tue, 29.09.26, 13:30–15:00, Room: Runder Saal

Chair: Jacobo de Uña-Álvarez

Two (few) subjects per variable for building continuous outcome prediction models

13:30-13:45

Presenter: Rebecca Whittle

Authors: Rebecca Whittle; Joie Ensor; Gary Collins; Richard Riley

Affiliations: University of Birmingham, Birmingham

Introduction

A 2015 study suggested that as few as two subjects per variable (2SPV) may be sufficient for fitting linear regression models for a continuous outcome [1]. Though originally focused on unbiased predictor effects, the 2SPV 'rule of thumb' has since been used to justify sample sizes in prediction model development this may be inappropriate as prediction modelling aims to optimise predictive performance. An alternative approach by Riley et al. proposes a four-criterion sample size targeting minimal overfitting, low optimism in R2, and precise estimation of the mean outcome and residual standard deviation (SD) [2]. We aimed to compare the performance of prediction models developed using sample sizes based on the Riley criteria and the 2SPV rule.

Methods

We conducted a simulation study comparing multivariable linear regression models using: (1) the 2SPV rule or (2) the minimum sample size from the Riley criteria. Data were generated from a super-population of 1,000,000 individuals with a continuous outcome (mean 120, SD 15). Included predictor parameters ranged from 2-20 and anticipated R2 from 0.10-0.90. For each scenario, 1000 samples were drawn and models fitted. We assessed shrinkage $\geq$0.90, optimism in R2 $\leq$0.05, and $\leq$10% margin-of-error for the mean outcome and residual SD.

Results

The 2SPV rule produced extremely small datasets (n=4-40) leading to highly overfitted models with poor performance. Shrinkage was typically >10% and optimism in R2>0.05. Precision targets were also frequently unmet, resulting in large uncertainty in predictions. In contrast, Riley-based sample sizes were larger and resulted in substantially less overfitting and better precision. The precision and minimal overfitting targets were attained on average, corresponding to an assurance probability of ~50% that a single dataset would achieve the targets or better whereas the 2SPV rule had very low assurance probabilities.

Discussion

Sample sizes from the 2SPV rule are not recommended for developing prediction models as they fail to optimise predictive performance and lead to very imprecise predictions. The Riley criteria addresses this on average, though with ~50% assurance of meeting targets. We describe how a generalised simulation-based framework allows researchers to achieve higher assurance probabilities [3].

1. doi:10.1016/j.jclinepi.2014.12.014

2. doi:10.1002/sim.7993

3. doi.org/10.48550/arXiv.2504.18730

Sample Size Determination for Developing Risk Prediction Models with Clustered Binary Data

13:45-14:00

Presenter: Izzati Izyani Japar

Authors: Izzati Izyani Japar [1]; Gareth Ambler [1]; Menelaos Pavlou [1]; Shafiqur Rahman [2]; Rumana Z Omar [1]

Affiliations: [1] University College London, London; [2] Institute of Statistical Research and Training (ISRT), University of Dhaka, Dhaka

Introduction

Clinical prediction models are often developed using multi-centre data, where within-cluster correlation must be accounted for. Random-effects models are recommended to enable cluster-specific predictions. A sufficiently large development sample size is essential to avoid overfitting and poor predictions in new data. Wynants (2015) recommended ensuring at least 10 events per variable (including the variance of the random intercept) for clustered binary data to limit bias in regression coefficients, rather than explicitly targeting predictive performance. Riley (2020) and Pavlou (2024) proposed sample size approaches for independent data that directly target predictive performance. We investigate whether these approaches are suitable for clustered binary data, and develop new formulae where required.

Methods

We conducted a full-factorial simulation to evaluate whether the Wynants method yields sample sizes sufficient for developing models with clustered data that have minimal overfitting and acceptable discrimination. Additionally, we assessed the Riley and Pavlou methods that were developed for independent data. Different simulation scenarios were investigated by varying multiple factors. Model performance was evaluated using the cluster-specific calibration slope (CS) and C-statistic, with target values specified for each. Using the simulation results, we developed our own sample size formulae using meta-models.

Results

As expected, the performance of the Wynants method varied depending on outcome prevalence and C-statistic. It resulted in overfitted models (CS0.2, with C-statistics sometimes >0.02 from the truth. The Pavlou and Riley methods targeting at CS=0.9 tended to underestimate the sample size for increasing number of clusters and clustering level. The Riley method worsened with increasing C-statistic and prevalence. For both approaches, the performance loss in terms of the C-statistic was small (≤0.015 of the truth). Our meta-model formulae worked well in providing the sample size required to achieve a target CS (e.g. 0.9) and target C-statistic (e.g. within 0.02 of the truth), by allowing inputs for the number of clusters and clustering level.

Discussion

Existing sample size approaches generally lead to models with inadequate predictive performance and may be unsuitable for clustered data. Two sample size formulae were proposed for clustered data, targeting calibration and discrimination respectively, that can be applied in many scenarios.

To transform or not to transform predictors for multivariable analysis - a benchmarking study

14:00-14:15

Presenter: Georg Heinze

Authors: Georg Heinze [1]; Daniele Giardiello [2]

Affiliations: [1] Medical University of Vienna, Wien; [2] Bicocca Center of Bioinformatics, Biostatistics and Bioimaging (B4 centre), School of Medicine and Surgery, University of Milano-Bicocca, Milan

Introduction

When fitting prediction models with continuous predictors, skewed or outlier-prone distributions can compromise robustness and precision in estimating predictor-outcome associations. Recently, an outcome-agnostic initial data analysis (IDA) has been proposed as a principled prerequisite to modeling (1). IDA provides insight into predictor distributions, and may motivate transforming predictors prior to modeling to reduce the influence of extreme values or simplify functional relationships. However, it remains unclear whether such outcome-agnostic transformations improve predictive performance or model simplicity.

Methods

We conducted a systematic comparison of outcome-agnostic transformation strategies across 50 publicly available datasets with varying outcome types and fixed predictor sets. We evaluated cubic root and (pseudo-)log transformations, assessing their ability to improve normality of predictors relative to no transformation. Effectiveness was quantified using changes in the Shapiro-Wilk statistic (squared Pearson correlation with normal deviates). Transformations were accepted if the statistic

improved by at least 0.2 (conservative strategy) or 0.05 (liberal strategy). Including the Shapiro-Wilk-optimal choice between cubic root and log transformations, as well as their combination, yielded six transformation strategies, which were compared to no transformation.

Regression models were fitted using linear predictor models with restricted cubic splines for continuous predictors. Complexities of the splines (1, 2, or 4 degrees of freedom) were selected via the Akaike Information Criterion (AIC). For comparison, we also fitted models assuming only linear functional forms. Predictive performance was compared using Akaike weights derived from AIC and summarized across datasets.

Results

On average, cubic root transformations achieved slightly higher Akaike weights than log-based approaches, and liberal strategies improved over conservative ones. However, the marginal benefit over a no-transformation approach was only small. In terms of the selected model complexity, conservative strategies generally resulted in simpler models than liberal approaches. Results were consistent when considering only linear functional forms. Despite these overall trends, in some data sets the choice of transformation strategy affected AIC dramatically and warranted further investigation.

Discussion

While there is probably no general benefit or harm of normalizing continuous predictors in generalized additive models, sensitivity to such pretransformations should be assessed when IDA reveals skewed predictor distributions.

(1) Heinze et al, BMC Medical Research Methodology, 2024. https://doi.org/10.1186/s12874-024-02294-3

A loss-based framework for coherent prediction of competing risks at prespecified time horizons

14:15-14:30

Presenter/Author: Hjalte Søberg Mikkelsen

Affiliations: Steno Diabetes Center Copenhagen, Copenhagen

Introduction

Accurate prediction of disease risk in the presence of competing risks remains a central challenge in clinical epidemiology. We target individualized predictions of the 5-year risk of type 2 diabetes (T2D) and simultaneously the 5-year risk of the competing risk of death. Common estimation methods, such as Fine-Gray regression and cause-specific Cox regression, target only one cause at a time and are not optimized for a given prediction time horizon.

Methods

We develop a loss-based estimation framework for the simultaneous prediction of multiple competing risks at a prespecified time horizon. Our approach provides a way to estimate a broad class of existing regression models for competing risks such that the resulting models are optimized for individualized risk prediction. Crucially, the framework enables seamless integration of flexible learner libraries and enables structural constraints during estimation. These may include coherence constraints that guarantee that the predicted risks respect the parameter space and domain-informed shape constraints (e.g., monotonicity in age). Estimation is achieved via constrained optimization of a differentiable loss function, such as the Brier score.

Results

We evaluate the proposed method in a large Danish registry-based cohort, where it provides individualized 5-year risk predictions for T2D and death without T2D. These predictions may be used for population-level monitoring of type 2 diabetes in Denmark, or to support clinical decision-making, for example in risk stratification of patients. The method demonstrates accurate and well-calibrated estimates of the cause-specific risks.

Discussion

This work introduces a flexible and unified framework for individualized prediction of cause-specific risks at a fixed time horizon. By directly optimizing a loss function at the target time point and enforcing structural constraints, the method yields coherent and clinically interpretable risk estimates. A central strength of the framework is its ability to incorporate flexible learner libraries, allowing it to bridge classical survival models with modern machine learning approaches. At the same time, the inclusion of domain-informed constraints ensures that increased flexibility does not compromise the validity or interpretability of the predictions. Ongoing work explores integration with advanced learners, including boosting and deep neural networks, to further enhance performance.

Integrating Landmarking and Competing Risks in Survival Analysis with Machine Learning Techniques

14:30-14:45

Presenter: Md Hasinur Rahaman Khan

Authors: Shirin Sultana [1, 2]; Md Hasinur Rahaman Khan [3]

Affiliations: [1] Bangladesh Institute of Governance and Management (BIGM), Dhaka; [2] Institute of Statistical Research and Training (ISRT), University of Dhaka, Dhaka; [3] University of Dhaka, Dhaka

Introduction

Modern survival analysis relies on dynamic survival prediction models that adapt to changing conditions using time-dependent covariates. Landmarking systematically assesses individuals at risk over

time, while mixed-model landmarking improves accuracy by incorporating longitudinal trajectories. However, competing risks, where multiple events influence outcomes, challenge traditional methods, leading to biased estimates and flawed clinical decisions. Integrating machine learning with mixed-model landmarking addresses these limitations, enabling more precise, data-driven predictions for real-world applications.

Methods

This study integrates landmarking and competing risks using various machine learning techniques. We assess the predictive performance of Random Survival Forest (RSF), LightGBM, XGBoost, and Bayesian Additive Regression Trees (BART) through simulation studies and real-world data from a multicenter Phase III breast cancer trial. The proposed mixed-model landmarking approaches are rigorously evaluated using several performance metrics, including Area Under the Curve (AUC), C-index, and Brier score.

Results

Across both simulation and empirical settings, mixed-model landmarking consistently outperforms standard landmarking in terms of calibration, discrimination, and predictive accuracy. Ignoring competing risks was shown to result in significant overestimation of event probabilities, underscoring the importance of appropriate model specification. RSF, in particular, demonstrated superior performance among the machine learning methods, yielding Brier scores as low as 0.021 and AUCs up to 0.883. These findings highlight the potential of combining longitudinal modeling with machine learning to capture dynamic risk profiles in complex survival settings.

Discussion

This study advances the field of dynamic survival analysis by integrating mixed-model landmarking with machine learning methods in the presence of competing risks. The resulting framework provides substantial improvements in predictive accuracy and clinical interpretability. Such approaches are particularly well-suited for modern healthcare data environments, characterized by high-dimensionality, temporal complexity, and non-linear interactions, positioning them as valuable tools for risk stratification and decision support in precision medicine.

Temporal Recalibration Using Pseudo-Observations in Prognostic Modelling: Addressing Informative Left Truncation

14:45-15:00

Presenter: Nickson Murunga

Authors: Nickson Murunga [1]; Sarah Booth [2]; Paul Lambert [3, 4]; Mark Rutherford [2]

Affiliations: [1] Biostatistics Research Group, Department of Population Health Sciences, University of Leicester, Liecester; [2] Biostatistics Research Group, Department of Population Health Sciences,

University of Leicester, Leicester; [3] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm; [4] Cancer Registry of Norway, Norwegian Institute of Public Health, Oslo

Introduction

In many time-to-event prognostic models, the Cox model is used, commonly assuming proportional effects over follow-up (1). When interest lies in prognosis at specific timepoints after diagnosis, Pseudo-Observations (POs) can be used to directly model covariate effects on survival at the timepoint of interest (2). A further issue in prognostic model development is the issue of calibration drift, where the predictions from prognostic models can become outdated when outcomes improve over time. We propose combining the PO approach and temporal recalibration to provide up-to-date failure probabilities for newly diagnosed patients (3). This requires defining a recent calendar period window and utilising delayed entry techniques. If entry depends on patient characteristics, informative left truncation may arise, potentially biasing POs (4). We therefore assess risk-stratified and weighted infinitesimal POs through simulation.

Methods

We performed a simulation study to assess the performance of POs methods under informative left truncation. Realistic simulation parameters were obtained using historical cancer registration data for colon cancer. Covariate distribution of age and stage were mimicked, and survival times generated from a Weibull model. Delayed entry was introduced through a recent 3-year period window. Five-year POs were estimated using the standard jackknife, risk stratified-stopped jackknife, and weighted infinitesimal stopped approaches. Performance was evaluated across varying sample sizes using bias, empirical standard error, and coverage probability. We further illustrate the approach using SEER colon cancer data observed failure probabilities for recently diagnosed patients were compared with predictions from standard and temporally recalibrated models using POs.

Results

The standard jackknife POs produces substantially biased estimates with poor coverage under informative left truncation. Risk-stratified and infinitesimal approaches produced negligible bias, appropriate empirical standard errors, and coverage close to 95% across a range of sample sizes. In SEER data, temporal recalibration using the proposed unbiased approaches showed improved calibration compared with a standard model.

Discussion

Care is needed when utilizing POs with temporal recalibration, as the standard approach produces biased estimates when delayed entry is covariate dependent. Risk-stratified and weighted infinitesimal approaches perform well under this scenario and can be utilized for POs to obtain up-to-date risk predictions for newly diagnosed patients.

Wed 30.09.2026

09:00–10:30

Dose Finding and Adaptive Trial Design

Wed, 30.09.26, 09:00–10:30, Room: K5

Chair: Babak Choodari-Oskooei

Design Informed Prior Calibration for BLRM and Emax Models in Oncology Phase I Dose Escalation Trials

09:00-09:15

Presenter: Lukas Andreas Widmer

Authors: Lukas Andreas Widmer [1]; Sebastian Weber [1]; Andrew Bean [2]

Affiliations: [1] Novartis Pharma AG, Basel; [2] Novartis Pharmaceuticals Corporation, East Hanover

Introduction

Model-based designs such as the Bayesian Logistic Regression Model (BLRM) with Escalation With Overdose Control (EWOC) are widely used in Phase I oncology trials. Up to corner cases, they have an equivalent functional form to sigmoid Emax models pervasively used in Pharmacometrics. Recently, several publications have highlighted a recurring issue when applying the BLRM: escalation can stall even when data suggest safety, e.g., when the posterior overdose probability at the next dose remains above the EWOC threshold despite observing no dose-limiting toxicity events (DLTs) at current or lower doses. Additionally, early extreme outcomes, such as a single DLT in a one-patient cohort, may have disproportionate impact on trial operating characteristics. These issues emphasize the need for carefully calibrated priors.

Methods

The starting dose and the planned escalation grid already encode a strong, asymmetric belief that implies a prior on the dose-DLT response curve: the starting dose is chosen such that it is likely safe, whereas uncertainty grows towards higher dose levels, justifying running a Phase I trial in the first place.

We evaluate prior specifications that formalize these assumptions using two parametrizations: the BLRM (log-intercept, log-slope) and Emax (ED33, h). Prior location, variance, and correlation are tuned using prior predictive checks against a library of "reasonable" escalation or dose-response trajectories. Calibration ensures that repeated 0/3 DLTs support escalation, isolated early DLTs do not prematurely halt the trial, and decisions for the first cohort qualitatively match simple designs such as 3+3. The design retains standard EWOC and interval-based overdose probability limits.

Results

We evaluated model behavior across scenarios varying curve steepness, MTD location, and early extreme outcomes. With calibrated priors for both model types, escalation stalls were eliminated, robustness to early DLTs improved, and early decisions aligned more closely with intended design behavior.

Discussion

It is imperative to choose a prior that reconciles implicit design beliefs in Phase I dose escalation trials when using model-based methods such the BLRM or Emax models to avoid recurring issues in practice. Under model misspecification, e.g., forgetting to model background treatment toxicity, more subtle differences between the models should also be considered.

Comparisons of intra-patient dose escalation approaches in phase I clinical trials

09:15-09:30

Presenter: Iphigenie Vannier

Authors: Iphigenie Vannier [1]; Sylvie Chevret [1]; Lucie Biard [2]

Affiliations: [1] APHP, Department of Biostatistics, Saint-Louis hospital, Paris; [2] ECSTRRA Team, IRSL, INSERM UMR1342, Université Paris Cité, Paris

Introduction

The primary objective of phase I oncology clinical trials is to determine the maximum tolerated dose (MTD) of a new drug, with toxicity serving as a surrogate for efficacy. Model-based approaches such as the continual reassessment method (CRM) [O'Quigley, 1990], have been shown to provide accurate MTD estimates, defined as a percentile of the dose-toxicity relationship. However, these dose escalation designs have two major limitations. Firstly, patients, especially those included at the beginning of the trial, may receive a subtherapeutic dose. Secondly, a minimum number of patients is required to obtain a reliable estimate of the MTD.

An interesting alternative is to extend the CRM to include intra-patient dose-escalation (IPDE). This approach could allow trial participants to access higher dose levels with greater expected efficacy, and improve MTD estimation by providing multiple observations per patient.

Our aim was to develop and evaluate the performance of various IPDE approaches in phase I clinical trials in terms of MTD identification, dose allocation, and safety.

Methods

We proposed and compared several designs for an extended standard CRM that incorporates IPDE [Guo, Liu, 2025]. We conducted a simulation study incorporating diverse scenarios with different shapes of dose-response curves, to assess the operating characteristics of each design. Various metrics were reported, including the percentage of correct selection, the average percentage of

patients treated at, below or above the MTD, the proportion of cycles in which a dose-limiting toxicity was observed, and an overdosing indicator, to evaluate the accuracy and safety of the approaches.

Results

The simulation study shows that the proposed designs efficiently identify the MTD. The IPDE increases the proportion of patients treated at the MTD while maintaining safety outcomes that are comparable to those of the standard CRM.

Discussion

We proposed and compared the performances of intra-patient dose escalation approaches with the CRM without IPDE. Our findings suggest that in various scenarios, IPDE can enhance dose allocation and MTD identification while maintaining stable safety metrics, emphasising its relevance for phase I trials.

Personalized Bayesian Phase I/II Trial Design Incorporating In-vitro Data: Application to Organoid and Cystic Fibrosis

09:30-09:45

Presenter: Yifan Mei

Authors: Yifan Mei [1, 2]; Sarah Zohar [3, 4, 5]; Kors van der Ent [2]; Danya Muilwijk [2]; Moreno Ursino [1, 3, 4, 5]

Affiliations: [1] Université Paris Cité, Paris; [2] University Medical Center Utrecht, Utrecht; [3] Inserm, Université Paris Cité, Inria, UMR 1346, HeKA, Paris; [4] Inria, Paris; [5] Inserm, Paris

Introduction

In vitro models are becoming increasingly important in precision medicine, particularly for rare diseases, where patients' variability is high but clinical trial sample sizes are limited. Existing phase I/II dose-finding designs provide limited statistical methods for incorporating such information into adaptive dose assignment, especially when in vitro readouts arise from complex assays indexed by an external driver or experimental factor, such as in organoids.

Methods

We develop a two stage joint Bayesian latent-variable framework for personalized phase I/II dose finding that corporates in vitro (patient's organoids) measurements, toxicity, and patient-specific efficacy. Patient-level latent variables capture unobserved treatment heterogeneity and link the in vitro measurement layer to an individualized efficacy model, allowing in vitro readouts and clinical outcomes to inform the same underlying sensitivity structure. Stage I uses toxicity-guided escalation to identify the maximum tolerated dose and a population-level admissible dose set. Stage II combines population-level and individualized admissible dose sets and applies patient-specific response-adaptive randomization under safety constraints. Gauss-Hermite quadrature is used to

marginalize latent random effects and make posterior computation tractable. The method, along with simpler models, was assessed through simulation studies inspired by forskolin-induced swelling readouts from intestinal organoids of patients with cystic fibrosis.

Results

Simulation studies spanning toxicity-dominant, efficacy-dominant, flat, and mountain-like dose-distribution scenarios show that individualized designs outperform traditional dose-finding designs, and that the proposed joint approach generally improves upon simpler plug-in strategies based on fixed in vitro summaries. On test sets, the joint design improves Top1 dose-selection accuracy by approximately 8 percentage points relative to plug-in methods and 7 percentage points relative to fixed-dose, while incurring limited additional safety cost.

Discussion

The proposed framework provides a robust approach for integrating in vitro response information into adaptive phase I/II dose finding. This method is effective at improving patient-specific dose stratification and admissible dose selection, but is limited in its ability to sharply distinguish between closely adjacent dose levels.

PK/PD-Integrated Bayesian Platform Design for Phase II Regimen Optimisation

09:45-10:00

Presenter: Axel Vuorinen

Authors: Axel Vuorinen [1]; Antoine Guillon [2, 3]; Emmanuelle Comets [4]; Moreno Ursino [1]

Affiliations: [1] Inserm, Université Paris Cité, Inria, UMR 1346, HeKA, Paris; [2] Inserm, Centre d'Étude des Pathologies Respiratoires (CEPR), UMR 1100, Université de Tours, Tours; [3] Service de Médecine Intensive Réanimation, CHRU de Tours, Tours; [4] Univ Rennes, Inserm, EHESP, Irset (Institut de recherche en santé, environnement et travail) - UMRS 1085, Rennes

Introduction

New approaches for early-phase dose-finding increasingly account for simultaneous toxicity and efficacy assessments. Incorporating pharmacokinetics/pharmacodynamics (PK/PD) measures into the modelling framework allows regimens to be compared based on drug exposure and biological activity, accounts for interpatient variability, and captures how dosing schedules with similar total dose affect drug exposure. Our objective was to develop a phase II adaptive platform trial design, to compare dosing regimens proposed for an investigational drug, that integrates PK/PD modelling into a Bayesian decision framework.

Methods

The PK/PD-informed Regimen Optimisation Platform (PROP) design incorporates joint population PK/PD modelling to link dosing regimens to drug exposure and biological activity, enabling patient- and population-level predictions that subsequently inform toxicity and efficacy model evaluations. We propose a time-to-event (TTE) model for acute and cumulative toxicities using PK with follow-up discretised into dosing cycles. A Bayesian model-averaged TTE approach combining two models, biomarker- and exposure-driven efficacy models, supports PK/PD-informed regimen optimisation under model uncertainty and enables assessment of the added value of activity data.

The design is evaluated in an influenza ICU setting, under the assumption that a prior phase I study has identified three dosing regimens. Toxicity is recorded per dosing cycle as time to first adverse event and efficacy is defined as time to ICU discharge.

Results

The PROP design was evaluated in a simulation study versus a dose-based platform (DBP) design using a cumulative model for toxicity and a Bayesian Cox model for efficacy. Both designs achieved a similar probability of correctly graduating the optimal regimen. PROP was more efficient, with earlier arm graduation, produced more accurate toxicity estimates, while slightly overestimated cumulative toxicity for the highest-risk regimen. DBP design tended to overestimate efficacy and underestimate early toxicity for high-risk regimens. Patient allocation patterns were broadly similar across designs.

Discussion

Integrating PK/PD modelling into an early-stage design improves the ability to discriminate among dosing regimens. Future investigations will involve additional biological and clinical models, and co-variates to identify patient subgroups and optimize regimen by subgroups.

Using GPC with multiple prioritized outcomes to analyze dose-ranging trials

10:00-10:15

Presenter: Lucie Biard

Authors: Sylvie Chevret [1]; Lucie Biard [2]

Affiliations: [1] Department of Biostatistics, Hôpital Saint-Louis, AP-HP, Paris; [2] Université Paris Cité, Paris

Introduction

'Generalized pairwise comparisons' (GPC), to compare patients' outcomes between independent treatment groups, have been shown promising to evaluate treatment effect, particularly as a trade-off between efficacy and toxicity in randomized clinical trials, as an alternative to composite end-points. Dose-ranging trials have recently gained interest in oncology where they have been proposed

for the dose optimization of agents of new therapeutic classes such as immunotherapies. Our objective was to evaluate GPC approaches for the comparison of efficacy and safety, specifically in multi-arm randomised clinical trials for dose optimization.

Methods

The statistical framework includes the presence of multiple prioritised outcomes, computation of GPC statistics between arms and family-wise error rate (FWER) control approaches (Bonferroni, Holm, Hochberg, and gatekeeping) to address the multiplicity issues with more than two arms. We performed a simulation study based on a previously published trial evaluating three dose levels of nivolumab in patients with renal cell carcinoma (NCT01354431). The operating characteristics of the GPC approach were primarily assessed considering a time-to-event efficacy endpoint with a binary toxicity endpoint, then to two time-to-event endpoints, with the LogRank test used as a benchmark. The impact of an additive threshold on the time-scale used to nuance the comparison between the outcomes of the paired observations, and censoring rates, were evaluated.

Results

Regardless of the sample size, FWER was controlled below the desired level of 0.05 by all multiplicity approaches, except for large values of the threshold in relation to the level of censoring. The gatekeeping hierarchical procedure improved power compared with the other multiplicity methods when the assumed test hierarchy was correct.

Discussion

The main interest of GPC lies in their ability to efficiently combine multiple endpoints, whereas using traditional composite endpoints may lead to a loss of information and an interpretation biased by emphasizing the first - rather than the worst - endpoint. GPC could be applied for the analysis of randomized dose-ranging clinical trials, combining efficacy and safety outcomes in a single and providing an assessment of the overall benefit consistent with a dose optimization aim. Multiplicity correction approaches can be applied to control FWER.

Optimal response-adaptive randomisation using Bayesian decision theory in group sequential designs

10:15-10:30

Presenter: Corey Voller

Authors: Corey Voller; Haiyan Zheng; Christopher Jennison

Affiliations: Department of Mathematical Sciences, University of Bath, Bath

Introduction

Response-adaptive randomisation (RAR) modifies allocation probabilities during a trial to favour better-performing treatments. In many Bayesian RAR implementations, allocation ratios are driven by plug-in summaries (e.g., posterior mean) of the treatment effect. This raises two practical questions: first, how much additional benefit is obtained by incorporating posterior uncertainty more fully and second, how to optimise allocations using the entire posterior distribution to minimise the number of subjects receiving the inferior treatment. We address these questions by proposing a unified loss-based decision-making framework for sequential trials.

Methods

We embed RAR into group sequential designs (GSDs), which aim to minimise an asymmetric loss that penalises information assigned to the inferior arm and escalates exponentially with the magnitude of the treatment effect. The rules for early stopping are held fixed with comparisons focusing on alternative allocation rules. Patient allocation in each stage is driven by an estimate of the treatment effect. As baselines, we use plug-in rules based on (i) the maximum likelihood estimate and (ii) the posterior mean. We develop two posterior-based alternatives. Method (iii) replaces plug-in estimates with an allocation ratio selected to minimise the posterior expected loss, integrating over the entire posterior distribution. Method (iv) extends (iii) by weighting the loss according to the conditional probability that the trial will stop at the next analysis given the current treatment effect estimate when selecting the next-stage allocation ratio. Thereby prioritising allocation ratios that matter most to the next decision. All scenarios are assessed by simulation under multiple prior specifications to determine design robustness, reporting realised loss at the final analysis and Monte Carlo standard errors.

Results

Simulation results show that early stopping drives most of the reductions in loss, while RAR delivers secondary but consistent benefits. Using information from the posterior and further optimisation yields incremental improvements and we provide a transparent way to quantify the benefit of different RAR scenarios.

Discussion

We provide practical recommendations for combining RAR with group sequential monitoring and for interpreting how allocation updating choices translate into realised loss. A further extension is to multi-arm multi-stage trials under sequential monitoring and multi-step look-ahead allocation rules.

Causal Inference with Survival and Competing Events

Wed, 30.09.26, 09:00–10:30, Room: N1

Chair: Susanne Strohmaier

The causal interpretation of acceleration factors

09:00-09:15

Presenter: Richard Post

Authors: Mari Brathovde [1]; Hein Putter [2]; Morten Valberg [1]; Richard Post [3]

Affiliations: [1] Oslo University Hospital, Oslo; [2] Leiden University Medical Center, Leiden; [3] Erasmus Medical Center, Rotterdam

Introduction

Effect measures in time-to-event analyses, particularly hazard ratios, are known to have a complicated and often misleading causal interpretation in the presence of unmeasured heterogeneity due to built-in selection bias from conditioning on survival. Accelerated failure time (AFT) models offer an alternative by parameterizing effects on the survival time scale through acceleration factors. These are often informally interpreted causally, yet a formal justification under frailty and treatment effect heterogeneity has been lacking. This paper aims to rigorously establish the causal interpretation of acceleration factors.

Methods

We define conditional and marginal causal acceleration factors within a potential outcomes framework and study their relationship to observed quantities. Under standard causal assumptions, including exchangeability and consistency, we derive identifiability results showing when the observed acceleration factor equals its causal counterpart. We further extend the framework to accommodate unmeasured frailty and treatment effect heterogeneity, and consider both time-invariant and time-varying settings. Simulation studies are conducted to compare acceleration factors with hazard-ratio based estimands.

Results

We show that the observed acceleration factor coincides with the conditional causal acceleration factor under exchangeability and consistency, even in the presence of unmeasured frailty and heterogeneous treatment effects. Unlike hazard ratios, acceleration factors avoid selection effects because they are defined on the survival scale rather than conditioning on survivors. Simulation results demonstrate that acceleration factors accurately recover the causal effect, whereas Cox model estimands are biased and sensitive to frailty and censoring. In settings with effect heterogeneity, the acceleration factor becomes time-varying and reflects a population-level summary that varies across quantiles of the outcome distribution.

Discussion

Acceleration factors provide a robust and interpretable causal estimand for time-to-event outcomes, overcoming key limitations of hazard-based measures. Their invariance to unmeasured frailty highlights a fundamental advantage over hazard ratios. However, in the presence of effect heterogeneity, interpretation requires care, as time-varying acceleration factors summarize complex mixtures of individual effects. These findings support broader use of AFT-based estimands in causal survival analysis while emphasizing the need for cautious interpretation in heterogeneous settings.

Causal treatment effect decompositions with time-to-event outcomes under competing events

09:15-09:30

Presenter: Mikko Valtanen

Authors: Mikko Valtanen [1, 2]; Tommi Härkänen [2]; Kari Auranen [1]

Affiliations: [1] University of Turku, Turku; [2] Finnish Institute for Health and Welfare, Helsinki

Introduction

Analyses of time-to-event outcomes in health research are often complicated by the presence of competing events that prevent the event of interest (target) from occurring and may obscure the interpretation of the apparent treatment effect.

Comprehensive characterisation of treatment effect requires distinguishing between different mechanisms by which the treatment operates. One approach is through the controlled direct effect, reflecting the treatment effect on the target event if the competing events were eliminated1. In addition, a two-way decomposition into direct and indirect effects has been proposed, based on the assumption that treatment comprises two distinct components, one affecting only the target and the other only the competing event2.

Methods

We propose a four-way decomposition of the treatment's effect on the target, characterising how it arises due to distinct mechanisms involving both the target and competing events. Based on a causal model, the decomposition relies on cross-world estimands reflecting counterfactual scenarios in which the treatment affects these events as if set to conflicting levels. We define the cross-world estimands through a process combining the indicators of each event under elimination of the other. We specify exchangeability and consistency assumptions under which the decomposition is identifiable from empirical observations.

Results

We illustrate how the components of the decomposition can be used to distinguish between different scenarios of how the treatment and the competing event conspire to produce the apparent treatment effect. We demonstrate the approach with two empirical datasets where the treatment either delays or accelerates the competing event.

Discussion

Our decomposition is analogous to a four-way mediation decomposition3 with the competing event taking the role of the mediator. The components may be combined to reflect estimands of particular substantive interest. The empirical expressions for the natural direct and indirect effect coincide with those for the separable direct and indirect effects.

1. Young et al. A causal framework for classical statistical estimands in failure-time settings with competing events. Stat. Med. 202039:1199-1236.

2. Stensrud et al. Separable Effects for Causal Inference in the Presence of Competing Events. J. Am. Stat. Assoc., 2022117(537):175-183.

3. VanderWeele. A Unification of Mediation and Interaction: A 4-Way Decomposition. Epidemiology 201425(5):749-761.

Estimating a treatment effect for survival when assignment to the target population is partially hidden

09:30-09:45

Presenter: Marilena Müller

Authors: Marilena Müller; Dominic Edelmann

Affiliations: German Cancer Research Center, Heidelberg

Introduction

We consider a two-arm randomized clinical trial in precision oncology with time-to-event endpoint. Patients in the control arm receive standard of care (SOC) treatment whereas patients in the experimental arm are offered personalized treatment, e.g. on the basis of molecular characterization of the disease. However, some patients in the experimental arm will not receive personalized treatment due to various causes Among others the main cause is that the necessary genetic property to carry out the personalized treatment is not given. These patients also receive SOC. The portion of people in the population this is affecting can be large and is unknown.

Classical Intention-to-treat analysis compares the outcomes of the two trial arms irrespective of the actual treatment. This is meaningful to evaluate the effect of a policy.

Methods

An alternative estimand, which may be interesting in practice, is the treatment effect in the population that would receive personalized treatment if randomized to the experimental arm. Direct estimation of this effect is not feasible, since information on which patients in the SOC arm would have received personalized treatment if randomized to the experimental arm is missing.

In particular, the survival distribution in the SOC arm is a mixture of distributions for patients who would and patients who would not potentially receive personalized treatment. We applied a non-parametric approach inspired by Patra and Sen [1] combined with isotonic regression to estimate the survival curves of both components in the mixture.

The estimated curves are compared using the restricted mean survival time, hypothesis testing is performed via bootstrap and other resampling methods. Alternatively, a semi-parametric approach is proposed.

Results

Power and Type I error of the methods are compared in detailed simulation studies.

This work presents a novel rigorous model for the analysis of treatments effects in the presence of mixtures or asymmetric trials.

Discussion

Practical guidelines on when this approach is necessary or appropriate are provided, considering other approaches to this problem.

[1] Patra, R. K. and Sen B. (2015), Estimation of a Two-component Mixture Model with Applications to Multiple Testing, Journal of the Royal Statistical Society, Vol. 78 (4), 869-893

Heterogeneous treatment effects on survival from randomised and observational data via Bayesian forests

09:45-10:00

Presenter: Tijn Jacobs

Authors: Tijn Jacobs [1]; Wessel van Wieringen [2]; Stéphanie van der Pas [1]

Affiliations: [1] Department of Mathematics, Vrije Universiteit Amsterdam, Amsterdam; [2] Department of Epidemiology and Data Science, Amsterdam UMC location Vrije Universiteit, Amsterdam

Introduction

Randomised controlled trials (RCTs) provide unbiased treatment effect estimates but are often too small to reliably detect treatment effect heterogeneity across patient subgroups. Observational databases offer larger and more diverse samples, yet may suffer from unmeasured confounding.

Data fusion methods that combine both sources can improve the precision of heterogeneous treatment effect (HTE) estimation, but existing approaches are largely limited to continuous or binary outcomes and rely on parametric or penalised regression models that may miss complex effect modification patterns.

Methods

We propose a Bayesian nonparametric model for estimating HTEs on survival outcomes by combining RCT and observational data. We adopt an accelerated failure time formulation and decompose the expected log-survival time into three components: a prognostic function capturing baseline risk, a treatment effect function representing the conditional average treatment effect, and a confounding function that captures bias due to unmeasured confounding in the observational data only. Each component is assigned a Bayesian Additive Regression Tree (BART) prior, allowing flexible estimation of non-linear effects and higher-order interactions without pre-specifying their form. To guard against over-borrowing from potentially incompatible observational data, we incorporate a meta-analytic prior that adaptively controls information sharing between data sources based on observed agreement. Causal estimands such as differences in restricted mean survival time and survival probability are obtained as posterior functionals of the estimated survival curves, with uncertainty quantified through the posterior distribution.

Results

In simulation studies mimicking realistic clinical settings, our method yields more precise HTE estimates compared to RCT-only analyses while maintaining low bias under varying degrees of unmeasured confounding in the observational data. We illustrate the approach on a breast cancer application combining trial and registry data.

Discussion

The proposed method offers a principled framework for leveraging real-world data alongside RCTs to improve detection and estimation of treatment effect heterogeneity in survival settings, while explicitly accounting for confounding bias and adaptively controlling the degree of borrowing. Limitations include the assumption that treatment effects are transportable across data sources conditional on observed covariates, and that identifiability of the confounding function relies on sufficient covariate overlap between the RCT and observational data.

A two-step causal inference approach for survival outcomes under time-varying latent confounding

10:00-10:15

Presenter: Yuan Liu

Authors: Yuan Liu [1]; Marta Fiocco [2, 3, 4]; Johannes Hm Merks [3, 5]; Marta Spreafico [2, 4]

Affiliations: [1] Leiden University, Leiden; [2] Mathematical Institute, Leiden University, Leiden; [3] Princess Máxima Center for Pediatric Oncology, Utrecht; [4] Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden; [5] Division of Imaging and Oncology, University Medical Center Utrecht, University of Utrecht, Utrecht

Introduction

Time-varying confounders pose a central challenge in causal inference, as they both influence and are influenced by prior treatments and outcomes. When observed longitudinal covariates have multiple categorical levels, the number of possible combinations can grow rapidly, producing sparse or emp ty data cells that violate the positivity assumption and complicate causal estimation. Addressing time-dependent confounding is even more challenging when confounders are latent, since the true disease state must be inferred from noisy data, potentially introducing misclassification bias.

Methods

We propose a two-step approach to estimate the causal effect of a time-varying exposure on survival outcomes in the presence of visit-specific latent confounders. First, we use latent Markov models to identify time-varying latent states while mitigating misclassification bias by estimating relative risks of experiencing one condition versus another over time, rather than assuming exact inference from observed data. Second, we estimate marginal structural models with different treatment histories via inverse probability weighting for treatment assignment and informative dropout. Average treatment effects (ATEs) are quantified in terms of differences in restricted mean survival times across alternative exposure strategies.

Results

The proposed approach was illustrated through a simulation study and an application to the effectiveness of chemotherapy treatment adjustments - dose reductions and cycle delays (yes/no) - in patients with Ewing sarcoma, a challenging setting due to the presence of chemotherapy-induced toxicities. Longitudinal chemotherapy administration data from the Euro Ewing 2012 trial were analyzed. Based on observed severity across multiple adverse events, we inferred cycle-specific relative risks of moderate/high latent overall toxicity (LOTox) during chemotherapy versus low LOTox. By incorporating these estimates as confounders, we assessed the ATEs of different treatment adjustment histories on event-free survival (EFS). Results did not indicate a clear detrimental effect on EFS, suggesting that toxicity-management adjustments may be implemented without evident harm to patients.

Discussion

This work presents a novel two-step causal inference approach for time-varying exposure with latent confounders. Future research should focus on establishing its theoretical properties, including consistency, efficiency, and robustness.

Funding: This study was supported by Stichting Kinderen Kankervrij (grant 458).

Acknowledgements: We gratefully acknowledge CRCTU and the EE2012 group for providing the data.

Doubly valid and doubly sharp sensitivity analysis to unobserved confounding for survival outcomes

10:15-10:30

Presenter: Jean-Baptiste Baitairian

Authors: Jean-Baptiste Baitairian [1, 2, 3]; Bernard Sebastien [2]; Rana Jreich [2]; Sandrine Katsahian [4, 3, 1]; Agathe Guilloux [1]

Affiliations: [1] HeKA (UMR 1346) INRIA/INSERM/Université Paris Cité, Paris; [2] Sanofi, Gentilly; [3] Université Paris Cité, Paris; [4] CIC-EC 1418 - Paris HEGP, Paris

Introduction

Time-to-event outcomes are central in oncology and rare diseases, with treatment effects often summarized by the survival function and the Restricted Mean Survival Time (RMST). Estimating these quantities from real-world data typically relies on the assumption of no unobserved confounding, an assumption that is rarely met in practice. Partial identification of treatment effects using sensitivity analyses therefore appears as a promising solution.

Methods

Under the Marginal Sensitivity Model (MSM) [4], we introduce new doubly valid and doubly sharp (DVDS) bounds for the difference of survival functions and difference in RMST by building upon two estimators of the survival function under informative censoring [1] and the DVDS bounds originally established for binary and continuous outcomes [2].

Results

We compared our methodology to the direct optimization method of Lee et al. [3] on simulated data and showed that our proposition not only yielded sharper, but also faster-to-compute bounds. This was also observed in a real data application on Right Heart Catheterization (RHC), where we reached a conclusion similar to Dorn et al. [2] about the negative effect of RHC on survival in intensive care units.

Discussion

The informative censoring assumption enables the introduction of the censoring function to correct for right censoring. However, we assumed the censoring function is independent from unobserved confounding to avoid an additional sensitivity analysis to the violation of informative censoring, a limit that should be addressed in future works.

[1] Chao Cheng et al. "Addressing extreme propensity scores in estimating counterfactual survival functions via the overlap weights". In: American journal of epidemiology 191.6 (2022), pp. 1140-1151.

[2] Jacob Dorn, Kevin Guo, and Nathan Kallus. "Doubly-valid/doubly-sharp sensitivity analysis for causal inference with unmeasured confounding". In: Journal of the American Statistical Association (2024), pp. 1-12.

[3] Seungjae Lee, Ji Hoon Park, and Woojoo Lee. "Sensitivity analysis for unmeasured confounding in estimating the difference in restricted mean survival time". In: Statistical Methods in Medical Research 33.11-12 (2024), pp. 1979-1992.

[4] Zhiqiang Tan. "A distributional approach for causal inference using propensity scores". In: Journal of the American Statistical Association 101.476 (2006), pp. 1619-1637.

Interoperable Data and Digital Evaluation

Wed, 30.09.26, 09:00–10:30, Room: N2

Chair: Christopher Gundler

AI-assisted semantic harmonization of laboratory values and units for interoperable real-world data integration

09:00-09:15

Presenter: Daniel Neumann

Authors: Daniel Neumann [1]; Alexander Strübing [1]; Florian Schmidt [1]; Fabio Aubele [2]; Tatjana Beppler [3]; Martin Federbusch [4]

Affiliations: [1] Institute for Medical Informatics, Statistics and Epidemiology, Leipzig University, Leipzig; [2] Medical Data Integration Center, LMU University Hospital, Ludwig-Maximilians-Universität München, München; [3] Institute of Medical Statistics, Computer and Data Sciences (IMSID), Jena University Hospital, Jena; [4] Department for Clinical AI and Translational Medicine, University of Leipzig Medical Center, Leipzig

Introduction

The reuse of routine laboratory data for multicenter studies and AI-driven analytics is limited by semantic heterogeneity, inconsistent unit encoding, and missing or context-dependent reference ranges. Although standards such as LOINC and UCUM exist, real-world implementation remains incomplete, introducing risks of misclassification and reduced analytical validity.

Methods

We developed an AI-assisted framework for semantic harmonization of laboratory data within distributed ETL pipelines. Instruction-guided large language models (LLMs), specifically GPT-OSS-20b-based architectures, generate candidate mappings between local laboratory parameters and LOINC concepts. The system integrates literature-aware retrieval by querying biomedical sources such as PubMed and text-based research APIs (e.g., NFDI4Health) to contextualize analytes, units, and expected value ranges.

Unit normalization is performed using UCUM APIs, including syntactic validation and dimensional consistency checks. Additionally, LOINC-defined primary and secondary units are used to verify mapping correctness. Where structured metadata is absent, reference ranges are inferred from literature and validated against empirical distributions. A context-aware plausibility layer incorporates diagnosis and medication information to assess whether observed values are clinically coherent. All mapping suggestions are reviewed by laboratory and IT experts prior to implementation.

Results

The framework was applied to all laboratory parameters at Ludwig-Maximilians-Universität München. It enabled systematic detection of unit inconsistencies, ambiguous mappings, and implausible values. The integration of literature-based evidence and clinical context improved mapping accuracy and identified discrepancies not detectable through standard approaches. The system reduced manual effort while ensuring traceable and reproducible mapping decisions.

Discussion

This approach combines AI-driven inference, formal terminologies, and expert validation to address a key bottleneck in real-world data integration. Embedding such engines directly into ETL pipelines and into the generation of FHIR resources would enable early-stage quality control at the interface to routine data. This is critical for scalable, high-quality real-world evidence and trustworthy AI applications.

Bridging Research and Care: Integrating Patient-Reported Outcome Measures from REDCap into Clinical Routine

09:15-09:30

Presenter: Egidia Cenko

Authors: Egidia Cenko [1]; Thomas Ganslandt [1, 2]; Jonathan M. Mang [1]

Affiliations: [1] Uniklinikum Erlangen, Erlangen; [2] Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen

Introduction

Patient-reported outcome measures (PROMs) provide essential insights into patients' symptoms, functional status, and health-related quality of life, supporting value-based and patient-centered care. Although widely collected in research platforms such as REDCap, PROMs are rarely integrated into clinical routine, limiting their clinical impact. Legal frameworks in the European Union, including General Data Protection Regulation and the Medical Device Regulation (MDR) further constrain data use, especially when automated score calculation or decision support is involved. This work presents a pragmatic and legally robust approach for transferring PROM data from REDCap into clinical routine, facilitating the use of research-collected PROMs in routine care.

Methods

An HL7v2-compliant interface was developed to transfer raw PROM responses from REDCap into synedra, an archive system that contributes to the electronic health record within the hospital information system. By omitting automated scoring, the system remains outside the scope of the MDR. Completed, untransferred records are periodically polled, and study pseudonyms are mapped to encounter identifiers. The system is implemented in Python, containerized with Docker, and deployed in Kubernetes. It is orchestrated via a cron job using Argo Workflows. Messages containing

Base64-encoded PDF files are securely transmitted to the communication server and transferred to the clinical system.

Results

Baseline assessments of various PROMs in the psychosomatic domain, with a focus on post-infectious and persistent somatic symptoms, were successfully transferred and made accessible as PDFs within the hospital information system. Logging and acknowledgment mechanisms ensured traceability. Initial feedback indicated improved accessibility and usability of PROM data in clinical workflows.

Discussion

Document-level PROM integration from research to routine systems is feasible without triggering MDR classification. This approach is transferable to other centers and lays the groundwork for future FHIR-based standardization and alignment with national initiatives, such as the MII core dataset for PROMs.

The full paper of this abstract will be published in the GMDS series "Studies in Health Technology and Informatics" with doi: 10.3233/SHTI260980, URL:https://dx.doi.org/10.3233/SHTI260980

AI-Powered Claims Categorization in openIMIS: Enhancing Efficiency for Medical Officers

09:30-09:45

Presenter: Dragos Dobre

Authors: Dragos Dobre [1, 2]; Simona Dobre [1, 2]

Affiliations: [1] Swiss TPH, Allschwil; [2] University of Basel, Basel

Introduction

Health insurance schemes in low- and middle-income countries face a critical operational bottleneck in claims adjudication. Nepal's Health Insurance Board (HIB) illustrates the challenge at scale: over 5.9 million claims filed across 780 health facilities for 3.8 million insurees are reviewed by only 20 claims officers - far below the estimated 300 required. Manual throughput is capped at roughly 1,600 claims per day against a daily demand of 30,000, creating delays that undermine access to care and scheme sustainability. Automated claims categorization through artificial intelligence offers a pathway to address this gap.

Methods

We developed an AI-based claims categorization module within openIMIS, the open-source health financing platform used in over 20 countries. Using historical claims data from Nepal HIB, we designed two model iterations. Version 1 applied an ensemble of XGBoost, Random Forest, and Extra Trees classifiers with majority voting on raw claim-level features. Version 2 addressed observed data drift by extending the feature set with temporally aggregated variables - including item frequency

per insuree over 7-, 30-, and 90-day windows - improving model stability over time. Both versions were evaluated on standard binary classification metrics. The module was integrated into openIMIS via a FHIR-compliant API, supporting two deployment architectures: a dedicated Claim-AI instance and an embedded configuration within existing openIMIS deployments.

Results

The Version 2 model achieved an accuracy of 98.65%, with precision of 80.3%, recall of 62.5%, and an F1-score of 70.3% on the Nepal HIB test dataset. Automated flagging significantly reduces the volume of claims requiring manual review, allowing medical officers to focus on high-risk cases.

Discussion

Ensemble ML methods combined with temporally engineered features can substantially reduce the adjudication burden in resource-constrained health insurance settings. FHIR-based integration ensures interoperability across openIMIS deployments. Future work will target improved recall, multi-class categorization, and cross-country validation. Designed as a global good, the module is available to all openIMIS implementers.

A Graphical Inference Approach for Clinical Decision Support with the InteKRator Toolbox

09:45-10:00

Presenter: Valentin Kulagin, Robert Stoffregen

Authors: Valentin Kulagin [1]; Robert Stoffregen [2]; Daan Apeldoorn [2, 1]

Affiliations: [1] IU International University of Applied Sciences, Düsseldorf; [2] Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI), University Medical Center of the Johannes Gutenberg University Mainz, Mainz

Introduction

This work presents a visualization approach for inferences from knowledge bases (KBs) learned with the InteKRator toolbox [1] for clinical decision support (CDS). While black-box methods (such as neural networks) lack transparency, the approach is based on explicit rules. A technical user interface (UI) has already been implemented in [2] (Figure 1). The work presented here focuses on the inference visualization.

Figure 1: Technical UI with a KB learned from a demo dataset.

Methods

The central contribution is a graphical approach [3] for InteKRator's inference mechanism. When submitting a query (e.g., a set of patient attributes), the inference is represented as a tree-like structure:

an input layer (e.g., for patient attributes) a rule layer (comprising the rules of a learned KB) an inference layer (providing the conclusions) Paths indicate the inputs/rule contributing to the conclusion (Figure 2).

Figure 2: Visualization of inference paths for a learned KB.

Results

The approach is demonstrated using a KB learned from a breast-cancer dataset [4] (Figure 2).

The visualization renders the inference process interpretable, the browser-based access increases the usability, and the modular tree-based approach also scales to more complex multi-step inferences as described in [5].

Discussion

The work bridges the gap between InteKRator's algorithmic capabilities and the practical needs of clinicians. By visualizing inferences as trees, recommendations can be easily comprehended - a prerequisite for team discussions and for increasing trust in CDS systems.

Limitations exist for large KBs. Future work may address these issues, e.g., through collapsible subtrees and filtering mechanisms.

References

Apeldoorn, D., Panholzer, T. Automated Creation of Expert Systems with the InteKRator Toolbox. Studies in Health Technology and Informatics, 2021283:46-55. DOI: https://www.doi.org/10.3233/SHTI210540 GitLab Repository of the Technical UI for the InteKRator Toolbox [Internet]. 2026 [cited 2026 Mar 27]. Available from: https://gitlab.rlp.net/intekrator/ui GitHub Repository of the Inference & Workflow UI for the InteKRator Toolbox [Internet]. 2026 [cited 2026 Mar 27]. Available from: https://github.com/Vale-Coder/InteKRator-UI Zwitter M, Soklic M. Breast Cancer [Internet]. UCI Machine Learning Repository 1988 [cited 2026 Mar 27]. Available from: https://doi.org/10.24432/C51P4M Thevapalan, A, Apeldoorn, D, et al. Comparison and Incorporation of Reasoning and Learning Approaches for Cancer Therapy Research. Studies in Health Technology and Informatics, 2023307:161-171.

When should the next clinical trial be conducted? A Bayesian borrowing approach for AI-based digital medical devices

10:00-10:15

Presenter: Caroline Lawless

Authors: Caroline Lawless [1, 2, 3]; Sandrine Boulet [1, 2, 3]; Moreno Ursino [1, 2, 3]; Rodolphe Thiebaut [4, 5, 6]; Sarah Zohar [1, 2, 3]

Affiliations: [1] Inserm, Paris; [2] Inria, Paris; [3] Université Paris Cité, Paris; [4] Université de Bordeaux, Bordeaux; [5] CHU de Bordeaux, Bordeaux; [6] Inria, Bordeaux

Introduction

Unlike traditional medical products such as drugs, AI-based digital medical devices (DMDs) evolve rapidly. Although a clinical trial may establish the efficacy of an initial version, these results may not remain valid for later versions. Repeating a full clinical trial at each software update is therefore impractical. Bayesian borrowing of information has been widely studied in settings such as basket and platform trials, where related evidence sources are combined while accounting for heterogeneity. However, existing borrowing methods do not address the sequential evaluation of evolving AI-based DMDs. The objective of this work is to develop an adaptive Bayesian framework to quantify discrepancies across device versions and guide decisions on when a new clinical trial is warranted.

Methods

We propose a Bayesian approach for evaluating evolving AI-based DMDs. Building on the temporal effects adjusted (TEA) prior framework, we develop a sequential Monte Carlo (SMC) algorithm in R that dynamically updates posterior distributions of efficacy across versions. In contrast to the original TEA formulation, where commensurability between historical and current data is specified a priori, our approach estimates this commensurability from the data. This allows information to be borrowed when appropriate and down-weighted when discrepancies arise. The method produces a sequential discrepancy measure that serves as a decision criterion for determining whether accumulated evidence supports launching a new clinical trial.

Results

Simulation studies considered scenarios reflecting different patterns of efficacy evolution across device versions, including stable performance, gradual drift, and abrupt departures from the initial version. The proposed method controls type I error while maintaining good power to detect clinically meaningful deviations. Compared with an approach analyzing each version independently without borrowing across versions, the proposed framework accumulates evidence more efficiently and identifies efficacy changes earlier, enabling earlier decisions on whether a new clinical trial should be conducted.

Discussion

Our Bayesian sequential borrowing approach provides a principled framework for monitoring the evolving efficacy of AI-based DMDs. It offers a data-driven criterion for determining when a new clinical trial is warranted while producing updated posterior estimates of efficacy for the current device version.

Harmonising evaluation and reimbursement of digital medical devices in oncology: A European classification grid

10:15-10:30

Presenter: Loubna Cadi

Authors: Loubna Cadi [1]; Sandrine Boulet [1]; Silvia Moler-Zapata [2]; Wendy Nieto-Gutierrez [2]; Louisa Stüwe [3]; Sarah Zohar [1]

Affiliations: [1] Inserm, Université Paris Cité, Inria, UMR 1346, HeKA, Paris; [2] Instituto Aragonés de Ciencias de la Salud, Zaragoza; [3] Digital Health Delegation, French Ministry of Health, Paris

Introduction

Cancer incidence in Europe is projected to rise by 20% by 2040. Digital Medical Devices (DMDs), including AI-driven diagnostics and digital therapeutics, offer pathways to optimise care. However, while the EU Medical Device Regulation (MDR) harmonises safety, market access in EU Member States remains fragmented. National frameworks (e.g., Germany's DiGA, France's PECAN) operate separately with different evidentiary requirements, despite ongoing efforts towards convergence. A disease-specific classification is necessary because oncology involves complex, often toxic, treatment regimens and requires intensive long-term follow-up in this context, DMDs are uniquely positioned to improve patient engagement, adherence, and quality of life. A significant barrier to cross-border scalability is the lack of a standardised taxonomy. This study, part of the eCAN Plus Joint Action - Enhancing Digital Capabilities of Cancer Centres in Europe, aims to develop a harmonized oncology-specific classification grid for DMDs, according to their primary intended medical purpose, to support Health Technology Assessment (HTA) and reimbursement.

Methods

A sequential four-step approach was employed: (1) Establishment of a scoping review (2015-2025) of European regulatory and HTA frameworks (2) construction of a comprehensive inventory of CE-marked oncology DMDs by cross-referencing EUDAMED and national reimbursement lists and (3) an inductive grid adapted from the generic Boers et al. EvalEU-DMD matrix that was piloted using the DMD inventory to ensure it captured oncology-specific nuances. (4) Finally, a structured questionnaire was developed for European experts to validate the grid's clarity, relevance, and utility for HTA evidence-grading.

Results

The landscape review confirmed significant heterogeneity in how clinical evidence is weighted across Member States. We identified 130 CE-marked oncology DMDs currently available in the EU. Mapping revealed that reimbursement status is highly inconsistent across the EU, with many tools reimbursed in only one or two Member States, reflecting both evidence-weighting gaps and a lack of specific oncology pathways. This mapping also showed that generic frameworks fail todistinguish critical oncology phases, such as multidisciplinary coordination and survivorship. Consequently, we developed the Onco-DMD Classification Grid, structured around 5 care phases (Prevention & Screening, Diagnosis & Staging, Treatment & Intervention, Monitoring & Patient Support and Clinical Coordination and System Support) and 11 clear functional sub-categories. The grid successfully categorises complex 'hybrid' tools by focusing on their primary medical purpose rather than underlying technology (AI vs. non-AI).

Discussion

The Onco-DMD grid provides a 'common language' for innovators and payers. By categorising devices by functional clinical use, it facilitates the definition of standardized evidence requirements across Europe. This harmonisation is essential for reducing the digital divide in cancer care and supporting the goals of Europe's Beating Cancer Plan.

Omics Discovery

Wed, 30.09.26, 09:00–10:30, Room: N3

Chair: Christoph Dieterich

A Python Framework for AI-Based Multiomics Integration

09:00-09:15

Presenter: Mireia Andueza

Authors: Mireia Andueza [1]; Pablo Villoslada-Blanco [1, 2]; Sergio Sabroso-Lasa [1, 2]; Evangelina López de Maturana [1, 2]; Núria Malats [1, 2]

Affiliations: [1] Spanish National Cancer Research Centre (CNIO), Madrid; [2] CIBERONC, Madrid

Introduction

The integration of multimodal data has become increasingly important for understanding complex biological systems and improving predictive modeling across diverse biomedical and life science applications. However, combining heterogeneous high-throughput datasets remains challenging due to their high dimensionality, variability, and technical differences. Artificial Intelligence (AI) methods offer powerful tools to address these challenges, yet standardized, reproducible, and user-friendly analytical workflows for multimodal integration remain limited.

Methods

We developed a Python package that provides an end-to-end framework for AI-based analysis of multiomics data. The package includes standardized preprocessing pipelines for heterogeneous datasets, such as batch effect correction, missing data imputation, and normalization. It incorporates multiple feature selection techniques and supports a range of Machine Learning and Deep Learning models for predictive analysis. Additionally, the framework implements several data fusion strategies, enabling systematic comparison of early, intermediate, and late integration approaches. The package also provides tools for model evaluation, including hold-out and cross-validation, as well as performance metrics, and methods for model interpretability and assessing potential biological or clinical relevance.

Results

The framework was evaluated using a multiomics liquid biopsy dataset (Nguyen et al. 2023) to explore the performance of different architectures and data integration strategies. Our results highlight the importance of systematically comparing multiple fusion approaches and model configurations to identify robust predictive models. In particular, data integration with intermediate fusion was the most promising strategy, yielding improved predictive performance compared to single-omics models. The framework enables reproducible implementation of these workflows and facilitates the

exploration of diverse AI strategies for multimodal analysis and will be available on GitHub for the scientific community.

Discussion

This work provides a flexible and accessible resource for researchers aiming to apply AI methods to multimodal data across various domains. By combining methodological guidance with a user-friendly software implementation, the framework promotes reproducibility and supports robust integrative analyses. Although validated in a specific application context, the approach is broadly applicable across a range of research areas involving multimodal data integration. Furthermore, its modular design enables the seamless incorporation of new models, either by the authors or by the broader scientific community.

Systematic comparison of statistical and machine learning methods for transcriptomic cell type classification in high-dimensional small-data settings

09:15-09:30

Presenter: Samuel Böhm

Authors: Samuel Böhm [1]; Mona Schoberth [1, 2]; Oleg Borisov [1]; Yong Li [1]; Gabriele Greve [1]; Bayram Edemir [3, 4]; Owen Woodward [5, 6]; Hyun Jun Jung [7, 8]; Frank Hutter [9]; Lukas Westermann [10]; Anna Köttgen [1, 2]; Pascal Schlosser [1, 2, 11]; Michael Köttgen [2, 10]; Stefan Haug [1]

Affiliations: [1] Institute of Epidemiology and Prevention, Faculty of Medicine and Medical Center - University of Freiburg, Freiburg im Breisgau; [2] CIBSS - Centre for Integrative Biological Signalling Studies, University of Freiburg, Freiburg im Breisgau; [3] Witten/Herdecke University - Institute of Physiology and Pathophysiology, Witten; [4] Department of Medicine, Hematology and Oncology, Martin Luther University Halle-Wittenberg, Halle (Saale); [5] Department of Pharmacology and Physiology, University of Maryland School of Medicine, Baltimore; [6] Maryland PKD Research and Clinical Core Center, University of Maryland School of Medicine, Baltimore; [7] Biomedical Research Institute, Kyungpook National University Hospital, Daegu; [8] Division of Nephrology, Department of Medicine, Johns Hopkins University School of Medicine, Baltimore; [9] Institut für Informatik, Albert-Ludwigs-Universität Freiburg, Freiburg im Breisgau; [10] Renal Division, Department of Medicine, Faculty of Medicine and Medical Center - University of Freiburg, Freiburg im Breisgau; [11] Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore

Introduction

In biological and medical research, defining cell type identity for cell culture - based in vitro experiments is essential, as a close match to the in vivo phenotype is critical for the validity of experimental findings. However, verifying cell line identities remains difficult. Matching bulk RNA-seq profiles from cell culture models to reference cell types offers a principled approach to address this problem, yet which methods are best suited for this classification task remains unexplored. Using kidney

as a model system, we compared classical statistical similarity measures against state-of-the-art machine learning classifiers for transcriptomic cell type classification.

Methods

Four pseudobulk reference profiles were constructed (two human, two murine) from kidney scRNA-seq datasets. Bulk RNA-seq samples were classified using three statistical similarity measures (Spearman correlation, Euclidean distance, Poisson distance) and three machine learning classifiers (Random Forest, XGBoost, and TabPFN). The impact of gene set choice on performance was evaluated across three sets: genome-wide expression, curated marker gene lists, and most variable genes. Validation followed a three-step framework of progressively increasing difficulty: within-reference pseudobulk classification, cross-reference generalizability, and classification of bulk RNA-seq from primary microdissected kidney tubules

Results

Rank-based Spearman correlation and TabPFN emerged as the most accurate (90.6% and 91.3% accuracy, respectively). Curated kidney marker gene lists showed a significant advantage over other selection strategies. Both methods correctly identified all microdissected kidnCorrelation Analysisey tubule segments and were robust against non-kidney negative controls. Notably, TabPFN - a transformer-based in context learner, pre-trained on synthetic datasets - provides posterior class probabilities, enabling uncertainty quantification unavailable with distance-based methods. We provide open-source implementations of both approaches to facilitate adoption across biomedical research communities.

Discussion

State-of-the-art probabilistic classifiers such as TabPFN match or exceed rank-correlation-based methods in accuracy while providing posterior distributions for principled uncertainty quantification. Informed feature selection via curated marker genes consistently outperformed data-driven alternatives, demonstrating that here domain knowledge effectively compensated for the high-dimensional, small-sample challenge. Our implementations and the methodological comparison generalize beyond kidney and should be tested in future work in other tissue contexts.

Adding layers of information to scRNA-seq data using pre-trained language models

09:30-09:45

Presenter: Sonia Maria Krißmer

Authors: Sonia Maria Krißmer [1]; Jonatan Menger [1]; Johan Rollin [2]; Tanja Vogel [2]; Harald Binder [1]; Maren Hackenberg [1]

Affiliations: [1] Institute of Medical Biometry and Statistics (IMBI), Faculty of Medicine and Medical Center, University of Freiburg, Freiburg; [2] Institute of Anatomy and Cell Biology, Department Molecular Embryology, Faculty of Medicine, University of Freiburg, Freiburg

Introduction

Pre-trained language models promise to enrich analyses of single-cell data with additional layers of information leveraging large text corpora. Yet, it is still unclear how to best align the complementary information with the primary quantitative data. Such an alignment could be particularly beneficial for adding further layers of information to low-dimensional representations of single-cell data obtained by artificial neural networks. For example, an information layer with textual descriptions on functional programs could be added to single-cell representations, that can be used for detection of disease associations.

Methods

To achieve such an alignment, we construct text-based, structurally parallel training datasets from both scRNA-seq data and biomedical literature. We transform quantitative single-cell data into a textual representation, a defined set of expression-ranked highly variable genes charaterising each cell, commonly called cell sentences. A tailored biomedical literature dataset is collected through an NCBI´s PubMed based search strategy. Utilizing a triplet-construction strategy allows for jointly training small language models with a Siamese-BERT architecture on both information sources emplyoing contrastive losses to learn a common, literature-enriched representation.

Results

Analyzing the contextualized single-cell embeddings, we showcase downstream applications including cell type functionality annotation, identification of functional programs and related marker genes, and exploration of disease relationships. In an exemplary application to developmental data, incorporating temporal metadata enables capturing transitions consistent with lineage trajectories.

The framework also provides systematic embedding-based evaluation strategies for language models in single-cell biology, including similarity tests with labels and metadata, annotation accuracy, and trajectory alignment to quantify how well models capture biologically meaningful relationships.

Discussion

By training small encoder-only language models jointly on single-cell and literature datasets we aim to augment classical analysis workflows of single-cell data, a complementary apporach to current research employing large language models for the primary analysis of scRNA-seq data. Our tailored approach allows for a targeted exploration of a cell-representation enriched with domain knowledge.

Mapping spatial cell-cell communication programs by tailoring chains of cells for transformer neural networks

09:45-10:00

Presenter: Niklas Brunn

Authors: Niklas Brunn [1]; Laia C. Guitart [1]; Kiana Farhadyar [1]; Camila L. Fullio [2]; Jakob Kailer [1]; Tanja Vogel [2]; Maren Hackenberg [1]; Harald Binder [1]

Affiliations: [1] Institute of Medical Biometry and Statistics (IMBI), Faculty of Medicine and Medical Center, University of Freiburg, Freiburg; [2] Institute of Anatomy and Cell Biology, Department Molecular Embryology, Faculty of Medicine, University of Freiburg, Freiburg

Introduction

Recent advances in spatial transcriptomics and computational modeling have enabled the study of cellular interactions in situ. However, existing methods quantify ligand-receptor-mediated cell-cell communication pairwise or between predefined cell groups, producing overlapping signals and offering limited ability to summarize concurrent interactions into communication programs while localizing spatial hotspots.

Methods

We introduce scCChain, a transformer-based framework that integrates ligand-receptor activity into spatially resolved communication programs and localizes hotspots at spot and single-cell resolution. scCChain first derives candidate programs using structured dimensionality reduction based on the boosting autoencoder approach. These programs are characterized by sparse loadings over ligand-receptor pairs and are encouraged by design to capture complementary communication signals. Next, scCChain samples program-specific communication chains by linking transcriptionally similar sender cells to candidate receivers via weighted random walks on a distance-informed cell graph, thereby borrowing signal from similar neighboring cells. A transformer model then scores the resulting chains to prioritize communication programs and pinpoint communication hotspots across the tissue.

Results

Applied to human breast cancer spatial transcriptomics data at spot and single-cell resolution, scCChain supports both exploratory discovery of communication programs and targeted analysis of user-specified ligand-receptor pairs. In whole-transcriptome spot-level data, it prioritizes a communication program enriched for pro-angiogenic signaling and localizes its activity to invasive tumor regions. In high-resolution imaging-based data with a limited gene panel, targeted analysis of CXCL12-CXCR4 highlights communication hotspots at cellular resolution, while inspection of transformer attention weights reveals characteristic sender-receiver cell type relationships and spatial signaling ranges.

Discussion

By modeling spatial cell-cell communication as chains rather than isolated pairwise events, scCChain captures coordinated communication programs and their spatial organization. This provides an interpretable framework for exploratory and targeted analysis across spatial transcriptomics platforms in heterogeneous tissues.

From Molecular Abundance to Interaction Networks: Biomarker Discovery in Thyroid MALDI-MSI Data

10:00-10:15

Presenter: Giulia Capitoli

Authors: Giulia Capitoli [1]; Marta Nesteruk [1]; Vasco Coelho [2]; Vincenzo L'Imperio [3]; Vanna Denti [4]

Affiliations: [1] Bicocca Bioinformatics Biostatistics and Bioimaging Centre - B4, Department of Medicine and Surgery, University of Milano-Bicocca, Monza; [2] Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milano; [3] Pathology Unit, Department of Medicine and Surgery, University of Milano-Bicocca, Fondazione IRCCS San Gerardo dei Tintori, Monza; [4] Proteomics and Metabolomics Unit, Department of Medicine and Surgery, University of Milano-Bicocca, Monza

Introduction

Thyroid nodules are a common clinical finding, yet preoperative diagnosis remains challenging due to high rates of indeterminate biopsy results, often leading to unnecessary surgery. Matrix-Assisted Laser Desorption Ionization Mass Spectrometry Imaging (MALDI-MSI) enables the identification of proteomic fingerprints directly from tissue sections, but its clinical application is limited by acquisition costs and the lack of a clear biological ground truth. While abundance-based methods identify discriminative molecular features, they do not capture interactions underlying disease mechanisms. This study proposes an integrative framework combining simulated and real-world data to evaluate both feature-level and network-level biomarkers in thyroid neoplasms.

Methods

Regularized statistical models, including Elastic Net and sparse Partial Least Squares Discriminant Analysis (sPLS-DA), were applied to identify discriminative molecules based on abundance in high-dimensional data. To account for molecular interactions, a Copula Graphical Model for heterogeneous mixed data was used to reconstruct conditional dependency networks for each diagnostic group, followed by a network rewiring analysis to quantify structural differences between conditions.

In parallel, an in silico simulation framework was developed to generate synthetic MALDI-MSI datasets under controlled scenarios, varying sample size, noise levels, marginal distributions, and interaction structures, to assess the behavior of different methodological approaches.

Results

Simulation results show that when classes differ mainly in marginal distributions, standard and penalized methods can effectively identify discriminative features and achieve accurate classification. However, as class distributions become more similar, noise increases, or discriminative information shifts from individual molecules to their interactions, these approaches lose sensitivity. In contrast, network-based models capture changes in dependency structures and remain informative under these conditions.

In the real dataset, the limited overlap between methods, including only seven shared molecules between Elastic Net and sPLS-DA, reflects the absence of a defined biological ground truth. Only two molecules (3%) were consistently identified by all approaches, while five (7%) were selected by at least three methods, reducing the feature space to a smaller subset of candidates.

Discussion

Abundance-based and network-based approaches highlight complementary aspects of the data. Their integration supports the identification of a restricted set of recurrent markers and provides a framework for interpreting high-dimensional MALDI-MSI data in thyroid neoplasms.

Strategies in Clinical Trials

Wed, 30.09.26, 09:00–10:30, Room: N4

Chair: Joost van Rosmalen

Open-source statistical reporting templates for recruitment time and participant count predictions at the design stage of clinical trials

09:00-09:15

Presenter: Katarina Zatkova

Authors: Katarina Zatkova; Philip Heesen; Manuela Ott; Malgorzata Roos

Affiliations: University of Zurich, Zurich

Introduction

Many medical studies fail to recruit enough participants within a specified timeframe, due to unrealistic recruitment planning at the design stage. Typically, study designers need to predict recruitment time until the required number of participants has been recruited, and the cumulative count of participants collected by a certain time. Statistical methods to make these predictions are rarely used in current practice and the availability of software tools for such predictions remains limited (Heesen and Roos 2024). Although Carter (2004) implemented statistical methods based on the stochastic Poisson process for making time and count predictions, the link to the source code in SAS is, however, obsolete and the methods are currently unavailable to the study designers, highlighting the need for reimplementation. Furthermore, the recruitment information needed in protocols varies across funding agencies, necessitating unification of the approach to design-stage recruitment analysis.

Methods

We use open-source R and Quarto software to reimplement, extend and disseminate the methods proposed by Carter (2004) in the form of open-source statistical reporting templates.

Results

Several open-source templates for time and count predictions address constant recruitment rate over time, the piecewise-constant recruitment rate and confidence considerations. The templates provide distributional predictions, estimates, plots and interpretations based on the medically justified recruitment rate, the required sample size and the planned study duration.

Discussion

The open-source templates will be available to study designers for realistic planning of recruitment time and participant count at the design stage of clinical studies. The template output addressing unified requirements in protocols could become a standard part of submissions to funding bodies. Consequently, more studies would reach their recruitment goals due to more realistic planning, and a higher number of studies would be successfully completed.

References

Heesen, P. and Roos, M. (2024). Freely accessible software for recruitment prediction and recruitment monitoring of clinical trials: a systematic review. Contemporary Clinical Trials Communications, 39:101298. https://doi.org/10.1016/j.conctc.2024.101298

Carter, R. E. (2004). Application of stochastic processes to participant recruitment in clinical trials. Controlled Clinical Trials 25 (5): 429-36. https://doi.org/10.1016/j.cct.2004.07.002

Phase III Probability of Success when Early Development is Based on a Putative Surrogate Endpoint: a Vaccine Efficacy Case Study

09:15-09:30

Presenter: Alessia Eletti

Authors: Alessia Eletti [1]; Andrea Callegaro [2]; Francesca Solmi [2]; Silvia Noirjean [1]; Valentino Conti [3]; Fabian Tibaldi [2]

Affiliations: [1] GSK Vaccines, Siena; [2] GSK Vaccines, Wavre; [3] GSK Vaccines Institute for Global Health, Siena

Introduction

As phase III trial costs and durations rise, pharmaceutical companies increasingly use quantitative methods to decide if a drug should progress beyond phase II. A key method is the Probability of Success (PoS) for phase III, calculated using the power function averaged across a treatment effect distribution estimated from phase II.

Methods

This paper explores PoS's role, particularly in moving from phase II trials with putative surrogate endpoints to phase III trials with clinical endpoints. Since the relationship between these endpoints is often unknown, expert input is necessary (prior elicitation). We propose the bivariate meta-analysis and a copula-based extension to characterize their relationship, using visual tools to simplify parameters elicitation. Specifically, we begin by eliciting the marginal distributions of the two quantities of interest. Then, to assist in eliciting the concordance parameter, we use the distribution of the treatment effect on the clinical endpoint conditional on the treatment effect on the putative surrogate. To support its use in practice, the proposed method is implemented in an R Shiny App.

Results

Our approach is illustrated in a case study on the development of a hypothetical prophylactic vaccine assumed to have similar characteristics as the COVID-19 vaccine, linking immunological and clinical endpoints utilizing the proposed copula-based method. It shows how historical data and expert opinions are integrated to define prior beliefs about treatment effects and their association, using an interactive graphical tool implemented in the form of an R Shiny App to aid expert elicitation of the correlation between immunological and clinical endpoints. The case study highlights that a larger absolute correlation leads to a greater conditional PoS and more informative Phase II results for de-risking Phase III.

Discussion

This work introduces a method to calculate and exploit the PoS for early phase drug development, particularly when transitioning from phase II trials with surrogate endpoints to phase III trials with clinical endpoints where their relationship is unclear and requires expert elicitation. The approach utilizes a copula-based bivariate meta-analysis with visual tools to aid the expert elicitation of endpoint relationships, offering advantages over existing methods by preserving marginal prior distributions and enabling conditional distribution visualization.

A Clinical Trial Framework for Evaluating a Digital Twin Medical Device: Application to Adaptive IL-7 Injection Protocols

09:30-09:45

Presenter: Théo René

Authors: Théo René [1]; Sandrine Boulet [2]; Moreno Ursino [3]; Mélanie Prague [4]

Affiliations: [1] Université de Bordeaux, Bordeaux; [2] Inria, Paris; [3] Inserm, Université Paris Cité, Inria, UMR 1346, HeKA, Paris; [4] Université de Bordeaux, Inserm, Bordeaux Population Health Research Center, U1219, Bordeaux

Introduction

While antiretroviral therapy successfully controls viral replication in most people living with HIV, some remain low immunological responders and fail to restore their CD4+ T-cell pool. Because administering Interleukin-7 (IL-7) injections induces CD4+ T cell proliferation, adaptive treatment protocols consisting of repeated injections have been proposed to minimize the time spent below a critical threshold (500 cells/μL). They can be viewed as Medical Digital Twins (MDTs), dynamically forecasting individual responses and guiding treatment decisions. However, such tools pose new methodological challenges for clinical evaluation, as traditional trial designs are often ill-suited to these evolving interventions. We investigate the suitability of a two-arm clinical trial design to assess whether an MDT outperforms a standard of care (SoC).

Methods

We propose a methodological framework for designing a two-arm non-inferiority clinical trial to evaluate an MDT-guided adaptive injection protocol against SoC. The framework focuses on two key aspects: the selection of clinically relevant endpoints and the definition of a final analysis that is independent of the MDT model, to avoid bias in treatment evaluation. As a primary endpoint, we consider the Time spent Above the critical Threshold (TAT). Since its estimation may be challenging with few CD4 measurements, we assess its relationship with a simpler surrogate endpoint: a single CD4 count at a fixed time after the last injection. We further examine the number of injections as a secondary endpoint. Trial operating characteristics are then evaluated through Monte Carlo simulations of synthetic patients generated from an existing mechanistic model of CD4 dynamics following IL-7 administration, with comparisons based on type I error and statistical power across alternative endpoint definitions.

Results

In our example, we found a correlation between primary and surrogate endpoints ($r=0.89$, $p<0.0001$) in SoC. Under adaptive dosing, the correlation remains robust but exhibits structured heterogeneity. Sample sizes are suggested accordingly, ensuring stringent control of Type I error while preserving high statistical power across all patient profiles.

Discussion

Our design provides a rigorous framework to assess whether an adaptive injection protocol improves patient health while reducing the number of injections. However, a similar methodology could be applied more broadly to the evaluation of other MDTs.

Assessing comparability in single-arm trials for rare diseases, with time-to-event endpoints

09:45-10:00

Presenter: Elia Gonzato

Authors: Elia Gonzato [1, 2, 3]; Romy van Voorst [1, 2]; Johannes Berkhof [3]; Marjo van der Knaap [1, 2, 4]; Rik van Eekelen [5]

Affiliations: [1] Department of Child Neurology, Emma's Children's Hospital, Amsterdam University Medical Center, Vrije Universiteit, Amsterdam; [2] Amsterdam Leukodystrophy Center, Amsterdam Neuroscience, Cellular & Molecular Mechanisms, Amsterdam; [3] Department of Epidemiology and Data Science, Amsterdam UMC location Vrije Universiteit, Amsterdam; [4] Department of Integrative Neurophysiology, Center for Neurogenomics and Cognitive Research, Vrije Universiteit, Amsterdam; [5] Amsterdam UMC, Amsterdam

Introduction

Randomised controlled trials (RCTs) are considered the gold standard for assessing the causal effect of an intervention compared to a control condition. However, RCTs are not always feasible due to practical or ethical constraints. RCTs are especially difficult to set up in pediatric settings because of the latter, and even more so if the disease is ultra-rare.

An alternative to RCTs is the single arm trial (SAT) design. Data for the control arm usually comprises historical controls from a disease registry. Although the SAT is considered more ethical and encourages trial participation, the evidence it yields is less robust as the design is prone to confounding and/or selection biases resulting from a lack of randomization and non-concurrency of the trial and control arms. An additional complication is that the start of follow-up (i.e. time zero) of historical controls has to be similar to the trial participants. To align timelines, it is the most straightforward to consider matching every treated participant to suitable control(s) and use the treated's time zero, thereby targeting the average treatment effect in the treated (ATT).

Methods

We investigated methodological challenges that trial investigators might run into when designing and conducting a SAT with time-to-event endpoints. Our motivating example is in Vanishing White Matter, where the time zero is defined as time of symptom onset and child participants could enroll in the SAT many years later.

Results

We show examples of timelines to explain the issues, discuss time-related biases such as immortal time bias and survivorship bias and argue that a priori-defined criteria for comparability must be defined by domain knowledge, as the characteristics and timelines of the 'ideal control patient' depend on the disease-specific characteristics of trial participants. (1)

Discussion

In the discussion we reflect on the influence of the 'matching order' that matching without replacement induces and mirror our recommendations to techniques and formulations from the target trial emulation framework and pharmacoepidemiology study designs.

References:

1. Suissa S, Azoulay L. Metformin and the risk of cancer: time-related biases in observational studies. Diabetes Care. 201235(12):2665-73.

A general framework for using two-stage individual participant data meta-analysis to predict individual treatment effects

10:00-10:15

Presenter: Marta Mainetti

Authors: Marta Mainetti [1]; Konstantina Chalkou [2]; Frederik Philipona [3]; Georgia Salanti [2]; Orestis Efthimiou [2]

Affiliations: [1] University of Bern, Bern; [2] Institute of Social and Preventive Medicine, University of Bern, Bern; [3] University of Bern, Schafisheim

Introduction

Individual participant data meta-analysis (IPD-MA) of randomized controlled trials can be used to predict individualized treatment effects (ITE). When data access is constrained, two-stage approaches are often required. Current ITE applications remain limited to regression-based models. The aim of this work is to introduce a general framework for two-stage IPD-MA to predict ITE for a continuous outcome, applicable to any type of statistical or machine learning model.

Methods

We outline different types of models that can be used at the first stage to conduct trial-specific analyses, including ordinary least squares (OLS) and penalized regression models, machine learning (ML), and Bayesian models. We then describe alternative second-stage methods for pooling results, depending on the first-stage model: (a) multivariate meta-analyses of regression coefficients for OLS (b) weighted average of predicted ITEs, applicable to all models, where weights are obtained by bootstrapping or from an auxiliary OLS model (c) multivariate Bayesian meta-analysis, for Bayesian regressions and (d) weighted mixing of MCMC samples for the Bayesian models. A simulation study assessed the performance of all possible combinations of first- and second-stage approaches under various scenarios.

Results

Model performance in simulations depended on the study-specific sample size, treatment effect heterogeneity, and the form of assumed effect modification. In scenarios with non-linear treatment-covariate interactions and large samples, ML outperformed regression methods. Conversely, in scenarios with linear effect modifications, regression models performed better. Among methods for pooling ML models, inverse weighting based on an auxiliary OLS model outperformed bootstrapping. We found no differences between frequentist and Bayesian regression. Shrinkage models provided modest improvements over unpenalized regression.

Discussion

Our framework for predicting patient-level treatment effects using two-stage IPD-MA can accommodate both statistical or machine learning models. The optimal method to use depends on the data generating mechanism and should be decided following internal validation.

Efficient Bayesian inference for nonparametric additive location-scale AFT models with a cure fraction

10:15-10:30

Presenter/Author: Philippe Lambert

Affiliations: Université de Liège, Liège; Université Catholique de Louvain, Louvain-la-Neuve

Introduction

Markov chain Monte Carlo (MCMC) methods are a cornerstone of Bayesian inference, valued for their reliability in exploring joint posterior distributions. However, in complex additive models, slow mixing and convergence issues can hinder efficiency, especially in interactive modeling or with large datasets. Many conditional posteriors for regression and penalized parameters in additive components can be effectively approximated by Gaussian distributions (Lambert and Gressani, 2023).

Methods

We focus on semi-parametric additive models for non-Gaussian censored data, where additive terms are represented as linear combinations of penalized B-splines (Marx and Eilers, 1998). Building on the work of Rue et al. (2009) and the INLA methodology, we employ Laplace approximations to the conditional posteriors of penalized parameters. This approach yields accurate, computationally efficient approximations of the multivariate posterior for regression and spline parameters, while enabling simultaneous selection of smoothing parameters.

Results

After presenting the general methodology and a simulation study, we demonstrate the application of Laplace P-spline models in flexible additive regression for interval- and right-censored data (Lambert, 2021). Special emphasis is placed on mixture cure models, combined with a flexible location-scale accelerated failure time (AFT) model for non-cured subjects. The methodology is illustrated using colon cancer data from a clinical trial.

Discussion

The method enables the flexible inclusion of covariates to model their joint effects on cure probability and multiple aspects of the time-to-event distribution. It is fast, robust, and efficient, and provides a relevant alternative to existing semiparametric and parametric approaches.

References

- Lambert, P. (2021). Fast Bayesian inference using Laplace approximations in non-parametric double additive location-scale models with right- and interval-censored data. Comput. Stat. Data Anal., 161, 107250.

- Lambert, P. and Gressani, O. (2023). Penalty parameter selection and asymmetry corrections to Laplace approximations in Bayesian P-splines models. Statistical Modelling, 23(5-6): 409-423.

- Marx, B.D., Eilers, P.H.C. (1998). Direct generalized additive modeling with penalized likelihood. Comput. Stat. Data Anal. 28 (2), 193-209.

- Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J. R. Stat. Soc. Ser. B, 71(2), 319-392.

Recurrent Events Modelling

Wed, 30.09.26, 09:00–10:30, Room: N5

Chair: Liesbeth de Wreede

Event-driven type design for clinical trials with recurrent events

09:00-09:15

Presenter: Jingwen Zhang

Authors: Jingwen Zhang [1]; Satoshi Hattori [2]

Affiliations: [1] Department of Biomedical Statistics, Graduate School of Medicine, The University of Osaka, Suita; [2] Department of Biomedical Statistics, Graduate School of Medicine and Integrated Frontier Research for Medical Science Division, Institute for Open and Transdisciplinary Research Initiatives (OTRI), The University of Osaka, Suita

Introduction

It is common in randomized clinical trials with the standard survival outcome to follow patients until a prespecified number of events have been observed, a type of trial known as the event-driven trial. The event-driven design ensures that the target power for a specified type 1 error rate is achieved to detect the target hazard ratio, regardless of the specification of other quantities. For recurrent event outcomes, in the absence of within-subject correlation among events, a similar event-driven design can be employed. On the other hand, in the presence of the within-subject correlation, one needs to model the correlation among recurrent events in evaluating power and setting the sample size. However, information useful in modeling the within-subject correlation is limited at the design stage. Failing to consider the correlation properly may lead to underpowered studies.

Methods

We propose an event-driven type design for recurrent event outcomes based on the semiparametric marginal rates model. Instead of determining the analysis timing by a target number of events fixed at the design stage, we continuously monitor a blinded estimator of the robust sandwich variance. The final analysis is conducted when the monitored variance reaches the threshold corresponding to the desired power for a prespecified treatment effect. This design aims to preserve the target power without requiring the correct specification of within-subject dependence or other nuisance quantities.

Results

We evaluated operating characteristics through simulation and illustrated the method using chronic granulomatous disease trial data. Across simulation settings with correctly specified planning assumptions, the proposed method achieved empirical power close to the target 80% without severe inflation of the type 1 error rate. Under misspecification of planning assumptions, the proposed

method preserved target power by adaptively extending follow-up. In the real-data illustration, the proposed method identified an analysis time slightly earlier than the conventional event-driven design while yielding a similar treatment effect estimate.

Discussion

The proposed method provides a practical event-driven type design for recurrent-event trials that explicitly accommodates within-subject dependence and reduces sensitivity to nuisance-parameter misspecification. The method may, however, prolong trial duration if the initial sample size is set too small, so careful planning remains important.

A study of the patient weighted while alive estimand for recurrent events with competing risks

09:15-09:30

Presenter: Maja Pohar Perme

Authors: Maja Pohar Perme [1]; Per Kragh Andersen [2]; Henrik Ravn [3]

Affiliations: [1] Institute for Biostatistics and Medical Informatics University of Ljubljana, Ljubljana; [2] Section of Biostatistics, Department of Public Health, University of Copenhagen, Copenhagen; [3] Novo-Nordisk, Bagsværd

Introduction

Chronic diseases are often characterized through the repeated occurrence of unfavorable events, for example hospitalizations, but the risk of death may also not be negligible. This implies that we are observing two processes and need to define a clinically interpretable estimand of interest.

Methods

In this talk we present the patient weighted while alive estimand, which represents the average event rate per year lived. We study non-parametric estimation and regression methods for this estimand, the latter using either pseudo-values or inverse probability of censoring weighted estimating equations.

Results

The properties of these methods are assessed through simulation studies, and illustrated with a practical application using data from a type 2 diabetes trial. We find that the patient weighted while alive estimand is less sensitive to differences in terminal event rates than the average number of recurrent events, and that both regression approaches are feasible.

Discussion

The while alive estimand has a simple and interesting interpretation, we present its advantages and disadvantages and give guidelines for its practical use.

Non-parametric estimation of excess recurrent events

09:30-09:45

This is one of this year's ISCB StCA award-winning abstracts.

Presenter: Jonatan Hedberg

Authors: Jonatan Hedberg [1]; Ola Hössjer [2]; Therese Andersson [1]; Elisavet Syriopoulou [1]

Affiliations: [1] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm; [2] Department of Mathematics, Stockholm University, Stockholm

Introduction

Recurrent events, e.g. hospitalisations, occur frequently in epidemiological research. Interest often lies in understanding how a specific disease impacts the number of events an individual is expected to experience. Quantifying the number of excess recurrent events attributable to a disease is therefore of high interest. Although this is a challenging task, the relative survival framework can be extended to estimate such a measure. In this approach, the observed rates among patients are compared to general population rates to obtain a measure for the excess rates due to the disease.

Methods

We propose a new non-parametric estimator for the excess number of recurrent events attributable to a disease, while taking the competing risks of death due to the disease (net survival) and other causes (background survival) into account. This approach extends the estimator by Ghosh & Lin [1] to the setting of excess recurrent events and relative survival. Estimation can be done by plugging in estimates of the excess recurrent events rate, an estimate of the net survival and the population survival that is known from life tables. To estimate the excess recurrent events rate and the net survival, data from the general population for the events of interest are needed. Covariance estimation can be done either using a closed-form expression or by usage of the bootstrap.

Results

The involved stochastic processes in the estimator are all either monotone or (weighted) empirical processes, and the associated mappings are compactly differentiable. Therefore, using empirical process theory and the functional delta method we show that the estimator converges weakly to a zero mean Gaussian process. We also show that the covariance estimator is consistent. This allows for the creation of confidence intervals using quantiles from the normal distribution. Finally, we present findings from a simulation study and illustrate the method using Swedish registry data on cancer patients.

Discussion

We propose a non-parametric estimator for the cumulative number of excess recurrent events associated with a particular disease, taking mortality into account and show how its covariance can be estimated.

[1] Ghosh & Lin. 2000. 'Nonparametric Analysis of Recurrent Events and Death'. Biometrics

---

Efficient Ghosh-Lin Regression for Recurrent Events in Presence of Terminal Events: Methods and Clinical Applications.

09:45-10:00

Presenter: Giuliana Cortese

Authors: Giuliana Cortese [1]; Thomas Scheike [2]

Affiliations: [1] Department of Statistical Sciences, University of Padova, Padova; [2] University of Copenhagen, Copenhagen

Introduction

Marginal mean regression models for recurrent events with terminal events, such as the Ghosh-Lin model, are widely used due to their flexibility and minimal assumptions on dependence structures. However, standard inverse probability of censoring weighting (IPCW) estimators are often inefficient and sensitive to censoring model misspecification. We propose a semiparametrically efficient augmentation framework that improves both efficiency and robustness.

Methods

We derive augmented IPCW estimators based on influence function decomposition and semiparametric efficiency theory. The optimal augmentation term involves conditional expectations of the full-data score, which we approximate via a flexible dynamic regression augmentation with time-varying coefficients. This approach avoids restrictive parametric assumptions (frailty models) while retaining computational tractability. We evaluate performance through simulations under multiple dependence structures, including two targeted scenarios: (I) a low action setting with reduced event intensity (ii) misspecified censoring distributions with covariate-dependent censoring. We further establish robustness properties in randomized clinical trial settings under model misspecification. The method is illustrated in an application to recurrent catheter-related bloodstream infections in patients with chronic intestinal failure.

Results

The proposed estimators achieve substantial variance reduction relative to standard IPCW Ghosh-Lin estimators across all settings, with the largest gains observed for dynamic augmentation. Efficiency improvements are consistent across dependence structures and censoring levels. In low-action settings, estimators remain stable with improved performance. Under misspecified censoring, standard estimators exhibit marked bias, whereas dynamic augmentation shows robustness with

minimal bias. In the clinical application, augmented estimators yielded more precise inference while maintaining consistent effect estimates. In RCT settings, the estimator consistently recovers null treatment effects under mild conditions, ensuring valid inference despite working model misspecification.

Discussion

This work provides a practical route to improve inference in regression modeling of recurrent events data with terminal events in any setting. The proposed dynamic augmentation achieves variance reduction regardless of model correctness and mitigates bias under censoring misspecification, thus providing properties of robustness. Its general adaptability and theoretical results make it well-suited for many biostatistical applications, including randomized clinical trials and complex survival data with several covariates.

References: Ghosh & Lin (2002) Cortese & Scheike (2022, 2026) Tsiatis (2006).

A Copula-Based Semi-Competing Risks Model for Recurrent and Terminal Events in presence of Cure Fraction and its Efficient Estimation via Stochastic Optimization

10:00-10:15

This is one of this year's ISCB StCA award-winning abstracts.

Presenter: Nasrin Sultana

Authors: Nasrin Sultana [1, 2]; Md Hasinur Rahaman Khan [1]; Moudud Alam [3]

Affiliations: [1] Institute of Statistical Research and Training (ISRT), University of Dhaka, Dhaka; [2] Gazipur Agricultural University, Gazipur; [3] School of Information and Engineering, Dalarna University, Borlänge

Introduction

In biomedical studies, the semi-competing risk data are very often observed where a non-terminal event, e.g. disease recurrence or being cured, may be censored by a terminal event such as death. Conversely, a terminal event can also cause a censoring of the non-terminal events. The non-cured patients' recurrence of the disease may be influenced by some observed and unobserved factors, or random effects or frailties. The frailty induces dependence between recurrent times but, the recurrent time and the terminal event time can also be correlated, giving raise to dependent censoring. The situation can further worsen by presence of (latent) cure status.

Methods

To address these complexities, we propose a flexible frailty mixed binary-continuous tri-variate copula model that simultaneously captures the dependence between a binary cure indicator and continuous event times for recurrent and terminal events. The presence of frailty and complex dependence structure induced by the copula made the Likelihood based inference computationally challenging.

To overcome these difficulties, we develop an efficient estimation procedure based on Hierarchical Model Stochastic Gradient Descent (HMSGD) algorithm. The method combines Monte Carlo sampling of latent frailties from their posterior distribution with stochastic gradient-based optimization which avoids high-dimensional integration and Hessian computation. Simulation studies demonstrate that the HMSGD estimator achieves ignorable bias and small RMSE for the parameter estimates. The HMSGD approach was also found to be computationally efficient and it exhibited stable convergence even with relatively small Monte Carlo sample sizes.

Results

Application of the proposed models are demonstrated using two sets of secondary data on post-surgery hospital readmission for colorectal cancer recurrence and hospital readmission of head neck cancer patients after radiotherapy as well as simulation study.

Discussion

The proposed method provides a practical and flexible tool for analyzing semi-competing risks data with a complex dependence structure due to recurrence, terminal events, and cure fraction. The bias is generally small, and the RMSE decreases with increasing sample size in simultion studies. The empirical standard deviations are consistent with the variability observed in the estimates, indicating stable estimation. Coverage probabilities are close to the nominal level for most parameters, suggesting that the inference procedure performs adequately.

Comparative benefit-risk assessment of treatments for recurrent and terminal events using a weighted Andersen-Gill model on ordering scores

10:15-10:30

Presenter/Author: Ákos Ferenc Pap

Affiliations: Bayer AG, Wuppertal

Introduction

Comparing efficacy and safety of a new treatment versus control often uses a single composite metric such as the win ratio (Redfors et al., 2020) for time-to-event outcomes of varying severity. A modified win ratio for recurrent and terminal events is proposed, estimated using weights based on event severity.

Methods

Follmann et al. (2020) proposed estimating the win ratio for ranked events by constructing ordering scores. For each patient the worst event is selected and the ordering score equals the time to that event or censoring plus a rank-based multiple of the study maximum follow-up, with rank 0 for the worst event. A Cox model is fitted to these ordering scores using a multiple-interval-censoring structure so each patient contributes one or more records depending on event severity the exponentiated treatment coefficient is the loss ratio and its reciprocal is the win ratio. An extension of the

method is presented that analyzes recurrent and terminal events while accounting for event severity via weighted Andersen-Gill model, applying the same weight to all records for a given event: the most severe event receives weight 1, less severe events receive weights

Results

The distribution and power of win ratios are presented for simulated recurrent and terminal efficacy and safety data with varying incidence rates and treatment effects, estimated using severity-weighted Andersen-Gill models (with ordering scores). These results are compared with win ratio estimates from equally weighted Andersen-Gill models and with cause-specific hazard ratios from Cox models for time-to-first-event.

Discussion

For assessing the overall treatment effect across multiple efficacy and safety events, the win ratio derived from a severity-weighted Andersen-Gill model using extended ordering scores can provide more relevant insight than an unweighted win ratio or the Cox model estimate based on time to first event.

References:

Redfors B et al. The win ratio approach for composite endpoints: practical guidance based on previous experience. Eur Heart J. 2020 Dec 741(46):4391-4399.

Follmann D et al. Analysis of ordered composite endpoints. Stat Med. 2020 Feb 2839(5):602-616.

Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample study. Ann Stat. 198210(4):1100-1120.

11:00–12:30

Clinical Prediction

Wed, 30.09.26, 11:00–12:30, Room: K2

Chair: Doranne Thomassen

LiD-GLM: Lipschitz-constrained Deep Generalized Linear Models

11:00-11:15

Presenter: Tom Splittgerber

Authors: Tom Splittgerber [1]; Marvin N. Wright [2]; Niklas Koenen [2]; Werner Brannath [1]

Affiliations: [1] Competence Center for Clinical Trials Bremen, Universität Bremen, Bremen; [2] Leibniz Institute for Prevention Research and Epidemiology - BIPS, Bremen

Introduction

Throughout various areas of research, modern machine learning (ML) methods have shown the potential to fit models of unprecedented goodness-of-fit in highly complex settings, given enough training data. However, in critical areas of research, which often rely on “small data”, the high complexity of ML methods, in particular of neural networks (NNs), limits their applicability twofold: limited amounts of data hinder training and their black-box nature prevents meaningful interpretation. Hybrid models which integrate ML methods into well-studied traditional statistical models are an intriguing approach to address these issues by improving flexibility while trying to retain interpretability and robustness.

Methods

We present a conditional distribution model that extends Generalized Linear Models (GLMs) by pre-transforming the input covariates with an invertible residual NN, allowing for flexible estimation of distributional parameters. Unlike previous approaches, we enforce a hard upper bound on the degree of nonlinearity of this network by constraining its Lipschitz constant. This enables a conscious, situation-dependent decision how far the hybrid model is allowed to deviate from the base GLM.

We further extend this model through a so-called Normalizing Flow operating on the target variable, allowing for a controlled relaxation of the distributional assumptions inherent to a GLM.

As our model learns the full conditional density, not just point predictions, it immediately models aleatoric uncertainty.

Results

We evaluate our model in several data applications, where it is initialized with the weights of a trained GLM. In settings where no benefit is gained from nonlinearity, our model is consequently equivalent to this GLM. In more complex applications, we empirically show that the goodness-of-fit of an unconstrained model can be reached for large allowed Lipschitz constants.

Discussion

Unlike previous approaches for hybrid models, which often only implicitly restrict ML components, the direct constraint of the Lipschitz constant forces LiD-GLM to rely on their linear component and prevents them from becoming equivalent to an unconstrained NN from an interpretational perspective.

Furthermore, the Lipschitz constant is a measure of the compromise between flexibility and interpretability and can give researchers insight on whether a given application necessitates complex ML methods.

Evaluating EMOO as a Metric-Balanced Ensemble Framework for Multi-Objective Optimization in Imbalanced Medical Data

11:15-11:30

Presenter: Maryam Moradpour

Authors: Maryam Moradpour [1, 2]; Zully Ritter [3, 2]; Anne-Christin Hauschild [4, 2]

Affiliations: [1] Institut für Predictive Deep Learning for Medicine and HealthcareJustus-Liebig-Universität Gießen, Gießen; [2] Medical Informatics Department, University Medical Center Göttingen, Göttingen; [3] University Medical Center Göttingen, Göttingen; [4] Institute for Predictive Deep Learning in Medicine and Healthcare, Gießen

Introduction: Class imbalance can hinder the reliable detection of clinically relevant outcomes in binary clinical prediction. In this study, the positive class was predefined as the clinically relevant minority outcome therefore, accuracy-only hyperparameter selection can favor the majority class and reduce sensitivity for that outcome. Methods: We evaluated the previously published Ensemble Multi-Objective Optimization (EMOO) framework as a hyperparameter-selection and ensembling method rather than a diagnostic model itself. EMOO uses the Non-dominated Sorting Genetic Algorithm II (NSGA-II) to search Random Forest hyperparameter configurations across accuracy, sensitivity, specificity, and F1-score, while additionally minimizing their standard deviation (STD) to discourage imbalanced metric performance. Final Pareto-optimal models are combined through Adaptive Boosting (AdaBoost). EMOO was compared with baseline Random Forest, the Synthetic Minority Over-sampling Technique (SMOTE), and the Multi-Objective Optimization Framework (MOOF) on three public binary clinical datasets with imbalance ratios of 1.16-2.11, using repeated 10×5-fold stratified cross-validation and Decision Curve Analysis (DCA). Results: EMOO achieved the highest

sensitivity and geometric mean across all datasets and the highest specificity on the Heart Failure Clinical Records and Mammographic Mass datasets. On the Pima Indians Diabetes dataset, its higher sensitivity was accompanied by lower specificity than baseline Random Forest and MOOF. DCA showed higher net benefit than treat-all and treat-none strategies across relevant threshold ranges. Discussion: EMOO provides a metric-balanced approach to selecting and ensembling classifier configurations for moderately imbalanced clinical prediction tasks. External validation on independent clinical cohorts is required before routine use. The EMOO package and evaluation code are available at github.com/HauschildLab/EMOO and github.com/HauschildLab/EMOO4IMD.

The full paper of this abstract will be published in the GMDS series "Studies in Health Technology and Informatics" with doi: 10.3233/SHTI261008, URL:https://dx.doi.org/10.3233/SHTI261008

Dynamic w-year survival prediction using gradient boosted landmark supermodels

11:30-11:45

Presenter: Achilleas Stamoulopoulos

Authors: Achilleas Stamoulopoulos [1]; Giorgos Bakoyannis [2]; Giota Touloumi [1]

Affiliations: [1] Department of Hygiene, Epidemiology and Medical Statistics, Medical School, National and Kapodistrian University of Athens, Greece, Athens; [2] Department of Biostatistics and Health Data Science, Indiana University Fairbanks School of Public Health, Indianapolis, IN, USA, Indianapolis

Introduction

Dynamic survival prediction enables individualized risk assessment for clinical outcomes using updated follow-up information. In people living with HIV (PLWH), although AIDS-related events have declined with effective antiretroviral therapy, non-AIDS events (e.g. cardiovascular disease) contribute substantially to morbidity and mortality. Landmark supermodels1 provide w-year dynamic risk predictions by fitting Cox-based models with delayed entry, time-updated covariates and landmark-dependent effects in stacked landmark datasets. However, predictive performance may deteriorate when the imposed model is misspecified or when complex higher-order interactions are present. We propose a gradient boosting-based landmark approach that flexibly captures nonlinearities and high-order interactions.

Methods

Using the landmark supermodel partial likelihood as the loss function, we implement an extreme gradient boosting (XGBoost) algorithm in R by supplying the gradient and Hessian of this partial log-likelihood to iteratively update a tree-based ensemble. The fitted model produces individual log-relative hazard risk scores which combined with the Breslow estimator of the baseline cumulative hazard, yield w-year survival probabilities. To allow for landmark-dependent effects, landmark time is included in the design matrix. Algorithm performance is evaluated through time-dependent area under the ROC curve (AUC) and Brier score in simulations incorporating high-order interactions,

time-varying and non-linear covariate effects. Comparisons made with Cox-based landmark supermodel including available covariates and their interactions with linear and quadratic landmark time. Finally, the method is applied to PLWH data from the AMACS study to predict 5-year cardiovascular disease risk.

Results

The proposed method outperformed Cox-based landmark supermodel across all scenarios in terms of predictive performance. In out-of-training data, the AUC was 0.80-0.92 and the Brier score 0.07-0.16 across a series of landmark times, compared with 0.62-0.84 and 0.10-0.34, respectively, for the Cox-based models. In the AMACS data (427 events over 79,733 person-years), the proposed method achieved cross-validated AUCs of 0.73-0.76 and sensitivities of 0.74-0.78, compared with 0.60-0.63 and 0.36-0.46, respectively, for the Cox-based landmark supermodel.

Discussion

The gradient-boosted landmark supermodel outperformed Cox-based model in scenarios with complex covariate effects and high-order interactions. Future work will examine the impact of model hyperparameters on predictive performance.

References:

1. VAN HOUWELINGEN, H.C. Dynamic Prediction by Landmarking in Event History Analysis. Scandinavian Journal of Statistics, 200734:70-85.

Latent Neural-ODE to enable precision dosing from sparse data

11:45-12:00

Presenter: Benjamin Maurel

Authors: Benjamin Maurel [1]; Moreno Ursino [1]; Jean-Baptiste Woillard [2, 3]; Sarah Zohar [1]; Agathe Guilloux [1]

Affiliations: [1] Inserm, Université Paris Cité, Inria, UMR 1346, HeKA, Paris; [2] Inserm P&T, U1248, Université de Limoges, Limoges; [3] CHU Limoges, Service de Pharmacologie, Toxicologie et Pharmacovigilance, Limoges

Introduction

Model-informed precision dosing requires accurate forecasting of drug exposure under alternative dose regimens to tailor treatments to individual patients. Continuous-time generative models, such as latent ordinary differential equations (latent ODEs), offer a data-driven alternative to rigid parametric models. However, predicting continuous exposure trajectories under new, un-administered doses traditionally requires heuristic interventions, like discontinuous state perturbations, which entangle the patient's underlying biology with the specific administered dose.

Methods

To overcome these limitations, we introduce a fully probabilistic generative framework. First, sparse clinical observations are encoded into a low-dimensional latent representation of the patient's baseline physiology. Second, to perform counterfactual dose extrapolation, a learned affine operator (Feature-wise Linear Modulation (FiLM)) maps this baseline representation into the target dose domain. This expresses the dose change as a smooth displacement in latent space. The framework is trained using an explicit bidirectional penalty (Bures-Wasserstein) to ensure latent displacements faithfully track the target dose distribution.

Results

We validated our framework on two independent datasets of renal transplant recipients (n1=178 n2=75). The model structured the latent space along hidden physiological axes autonomously. In a controlled simulation study, traditional parametric Non-Linear Mixed-Effects (NLME) estimators degraded under structural misspecifications, such as non-linear saturation kinetics. In contrast, our flexible, data-driven generative model successfully adapted to these deviations, avoiding predictive degradation and reducing the mean Root Mean Squared Percentage Error (RMSPE) by 8.1 percentage points compared to the clinical NLME baseline. On real-world data, the model achieved high predictive precision, yielding an external RMSPE of 10.45%. Clinically, our probabilistic latent ODE reduced the number of large prediction errors (> 20% relative error) from 7 cases in the clinical benchmark to 4 cases. Furthermore, during counterfactual dose extrapolation on a simulated virtual cohort, the model precision did not decrease significantly.

Discussion

By mapping discrete dose interventions to a bidirectionally regularized affine modulation operator, this approach resolves the discontinuous jump problem that has previously hindered continuous vector fields in pharmacokinetics. This probabilistic latent ODE provides a principled, differentiable framework for counterfactual clinical decision-making and personalized dose optimization without requiring rigid structural assumptions or heuristic workarounds

---

Developing useful prognostic models to support what-if questions

12:00-12:15

Presenter: Eleni Rapsomaniki

Authors: Eleni Rapsomaniki [1]; Awa Diop [2]

Affiliations: [1] AstraZeneca BioPharmaceuticals, Real-World Data Science and Analytics, Cambridge; [2] AstraZeneca BioPharmaceuticals, Real-World Data Science and Analytics, Mississauga

Introduction

Traditionally, work on prognostic models has prioritized predictive accuracy. Hence, models often include strongly predictive but non-causal variables, such as treatments and frequency of healthcare use as proxies for disease severity, and may even incorporate covariates potentially affected by reverse causality. These models may be useful at population level, but it is unclear if they be reliablly used for "what-if?" questions and for prioritization of interventions at the individual level.

We aimed to compare predictive and causal-leaning approaches inspired from recent literature for identifying patients at high risk of COPD progression. For illustration, we focus on identifying patients most likely to benefit from smoking cessation, in a setting with time-varying confounding, reverse causality, and treatment-related distortion of measurements.

Methods

We simulated a longitudinal COPD cohort to model disease progression including non-linear effects, interactions, unmeasured confounding, and severity-driven smoking cessation. Using baseline covariates to build the prognostic model, we compared a complex machine learning prognostic model, a parsimonious prognostic model based on more plausible causal predictors, an IPTW-weighted g-computation approach to estimate risk under smoking cessation, and an SCM-inspired structural equation model approach to propagate effects through mediators. We used causal SHAP as an additional aid to explain predicted benefit from the intervention-based model. We assumed smoking cessation reduces risk directly and through an FEV1 pathway, but increases risk through an eosinophil pathway. We compared methods using the expected absolute risk reduction in the highest-risk decile selected by each method. Methods will also be illustrated using COPD cohorts from UK and US electronic health records.

Results

In simulations, the IPTW-weighted g-computation approach and the SCM-inspired approach gave the best overall balance between discrimination and targeting of cessation benefit. The parsimonious model was more conservative, with weaker risk stratification and weaker benefit targeting. The accuracy-first machine learning model discriminated well but gave misleading estimates of intervention benefit, consistent with confounding and reverse causality.

Discussion

Pure prediction accuracy may not be enough for treatment prioritisation. Simple causal adjustment and structural modelling ideas may help, but they make strong untestable assumptions. Our initial exploration highlights the many challenges and some promising approaches that could be optimized further.

A bootstrap-based framework for external validation of rare-event prediction models under case-mix and prevalence shift

12:15-12:30

Presenter: Ester Rosa

Authors: Ester Rosa [1, 2, 3]; Mariaenrica Tinè [1]; Loc Carlo Bao [4]; Gloria Brigiari [5, 2, 3]; Stefania Lando [6, 2, 3, 4]; Laura Bonanno [4]; Dario Gregori [5]

Affiliations: [1] Unit of Biostatistics, Epidemiology and Public Health, Department of Cardiac, Thoracic, Vascular Sciences, and Public Health, University of Padova, Padova, Italy, Padova; [2] BIOSTAT-X Biostatistics & AI for Biomedical Discovery, Pediatric Research Institute (IRP) 'Città della Speranza', Padova, Italy, Padova; [3] PhD Program in Translation Specialistic Medicine "G.B. Morgagni", Curriculum "Biostatistics and Clinical Epidemiology", University of Padua, Italy, Padova; [4] Department of Surgery, Oncology and Gastroenterology (DiSCOG), University of Padova, Padova, Italy, Padova; [5] Unit of Biostatistics, Epidemiology and Public Health, Department of Cardiac, Thoracic, Vascular Sciences and Public Health, University of Padova, Padova; [6] Unit of Biostatistics, Epidemiology and Public Health, Department of Cardiac, Thoracic, Vascular Sciences and Public Health, University of Padova, Padova, Padova

Introduction

Machine-learning models are increasingly proposed for clinical use, but external validation is especially difficult in small real-world cohorts with rare outcomes and distribution shift. We externally validated a random-forest model for immune-related pneumonitis in advanced non-small cell lung cancer (NSCLC) and examined whether prevalence- and case-mix-adjusted resampling could improve the interpretation of transportability.

Methods

Predicted risks were generated in an independent external cohort without model refitting. Apparent performance was evaluated using area under the receiver operating characteristic curve (AUC), Brier score, log-loss, and calibration intercept and slope, explicitly distinguishing discrimination, overall performance, and calibration. Uncertainty was quantified via non-parametric bootstrap to account for sampling variability under rare-event conditions. We then conducted a transportability-oriented sensitivity analysis using weighted bootstrap resampling of the external cohort, targeting the event prevalence observed in the development dataset and aligning predictor distributions through iterative proportional fitting (raking), thereby approximating development-like case-mix conditions. Logistic and isotonic recalibration were assessed in bootstrap split-samples to evaluate model updating under distributional shift. Finally, simulation-based resampling was used to estimate the external sample size required to achieve a prespecified precision for the AUC, explicitly incorporating outcome rarity and variability across repeated sampling.

Results

In the observed cohort, discrimination was unstable and calibration was poor, indicating limited direct transportability. Under prevalence- and case-mix-adjusted resampling, mean discrimination improved to 0.73 (ΔAUC +0.30) and overall prediction error decreased (Brier score 0.08), suggesting that between-cohort differences materially contributed to the apparent performance loss. However, variability remained substantial across bootstrap samples, highlighting the intrinsic uncertainty

of rare-event validation.Precision analyses indicated that the available external sample was underpowered around 400 patients were required to achieve an AUC 95% confidence interval width below 0.20 in single-seed simulations, while 500-1500 patients yielded more robust precision across seeds.

Discussion

External validation of AI-based models in small-data oncology settings should distinguish observed performance from performance under explicit transportability assumptions and should quantify precision directly. We propose a general, data-efficient validation framework for rare-event prediction that separates sampling variability from structural lack of transportability. A practical framework combining bootstrap validation, covariate-shift adjustment, recalibration, and validation sample-size planning may provide more interpretable evidence before clinical deployment.

Basket Trials and Adaptive Borrowing

Wed, 30.09.26, 11:00–12:30, Room: K5

Chair: Meinhard Kieser

Bayesian Assurance with Calibrated Go Rules for Basket Trials: A Commensurate Prior Framework

11:00-11:15

Presenter: Niamh Fitzgerald

Authors: Niamh Fitzgerald [1, 2]; Adrian P. Mander [2]; Helen Mossop [3]; James Wason [4]

Affiliations: [1] Population Health Sciences Institute, Newcastle University, Newcastle upon Tyne; [2] GSK, UK, London; [3] Population Health Sciences Institute, Newcastle University, Newcastle Upon Tyne; [4] Population Health Sciences Institute, Newcastle University, UK, Newcastle upon Tyne

Introduction

Basket trials evaluate a single intervention across multiple patient subgroups simultaneously, offering efficiency gains in early-phase development. A central statistical challenge is balancing information borrowing across baskets against heterogeneity in treatment effects while preserving well-controlled operating characteristics. When a fixed posterior probability threshold is applied within a Go-any decision rule, the probability of incorrectly declaring at least one basket successful increases with sample size, and is particularly concerning under partial null configurations where borrowing from truly effective baskets can inflate posterior evidence in inactive baskets. This work develops a Bayesian assurance methodology for basket trials that addresses this alongside uncertainty in design inputs.

Methods

A three-zone decision framework partitions trial outcomes into Go, Consider, and No-Go regions, providing graduated operational recommendations for each basket. Treatment effects are modelled using a commensurate hierarchical prior for adaptive information borrowing. A posterior Go threshold is calibrated as a function of sample size to control false-positive rates, with per-basket thresholds adjusted to require stronger posterior evidence in baskets where prior support for a treatment effect is weaker. Sample size is selected via assurance, defined as the prior predictive probability of trial success, using ratio-preserving scaling followed by coordinate-descent optimisation. The methodology also supports interim assurance-based sample size re-estimation, enabling adaptive adjustment of basket-level allocations mid-trial

Results

Simulation studies across eleven design scenarios demonstrate that the calibrated design improves both false-positive control and sample size efficiency relative to fixed-threshold and no-borrowing alternatives. The framework is illustrated through the Obeticholic Acid for the Amelioration of Cognitive Symptoms trial, with prior sensitivity analyses confirming robustness of the recommended designs to moderate perturbations in outcome variability, treatment effect magnitude, and allocation assumptions.

Discussion

The proposed methodology provides a principled and practical approach to sample size determination in Bayesian basket trials, supporting unequal allocation ratios, heterogeneous prior beliefs, and graduated decision-making. Sensitivity to prior distribution underscores the importance of careful prior elicitation at the design stage. Future extensions include binary and time-to-event endpoints, utility-based decisions, and multi-arm platform settings.

A Power Prior Based Basket Trial Design for Unequal Sample Sizes

11:15-11:30

Presenter: Lukas Baumann

Authors: Sabrina Schmitt [1]; Lukas Baumann [2]

Affiliations: [1] University of Würzburg, Würzburg; [2] University of Heidelberg, Heidelberg

Introduction

Basket trials examine the efficacy of a single intervention in several patient subgroups, called baskets. They are mostly applied in oncology, where assignment to the baskets is based on matching medical characteristics. This can result in small and heterogeneous basket sample sizes. Several designs for the analysis of basket trials have been proposed in the literature that share information across baskets to increase power. Most designs utilise Bayesian methods, such as hierarchical modelling and model averaging. The recently proposed power prior design uses empirical Bayes methods to increase the computational efficacy compared to fully Bayesian designs. This design incorporates data from all baskets using a weighted likelihood that shares information according to the similarity of the baskets. However, if the sample sizes differ, there is a risk that the information from the small baskets will be overlaid by that from the large baskets.

Methods

We extend the power prior design by applying a weighting method, previously suggested for sharing information from historical data, that accounts for unequal sample sizes by limiting the amount of information shared between baskets. The new weights take the pairwise ratio of sample sizes per basket into account, such that the effective sample size that is shared per basket cannot exceed

the sample size of the basket of interest. Using a simulation study, we systematically compare the power prior design with previously suggested weights and the new information-limiting weighting method to other Bayesian basket designs.

Results

The results of the simulation study show that the new information-limiting weights improve the results of the original power prior design. In terms of the expected number of correct decisions, the improved power prior design performs slightly better than the competing designs in all sample size scenarios. In scenarios with some active and some inactive baskets, the inflation of the type 1 error rates is less severe than with unlimited sharing.

Discussion

The proposed basket trial design shows promising performance in the investigated scenarios and is computationally less expensive than fully Bayesian designs. It can therefore be considered for the analysis of basket trials with unequal sample sizes.

Limiting the Borrowing in Bayesian Basket Trials via Effective Sample Sizes

11:30-11:45

Presenter: Paul Thalmann

Authors: Paul Thalmann; Lukas Baumann; Meinhard Kieser

Affiliations: Institute of Medical Biometry, University of Heidelberg, Heidelberg

Introduction

Basket trials are type of master protocol often employed in oncology, enrolling patients not primarily based on their shared histology, but a shared biomarker. Differing histological subgroups form distinct 'baskets", enabling simultaneous efficacy evaluation across multiple histologies. During analysis, Bayesian basket trials leverage information borrowing across sufficiently similar baskets to mitigate the often low number of patients and resulting low power. While including many baskets may seem logistically attractive, the effect of basket number on operating characteristics remains understudied. We hypothesize that even moderate per-basket borrowing can cumulatively bias the analysis when a large number of baskets is considered.

Methods

We build on the Power Prior framework (Baumann et al., 2024) for flexible, transparent information borrowing. Borrowing is quantified using effective sample sizes (ESS) via the method of Chen et al. (2025). Two limiting mechanisms are compared: (1) downweighting all contributing baskets proportionally, and (2) restricting borrowing to the most similar baskets. A simulation study systematically

evaluates both mechanisms against unlimited borrowing across scenarios, varying, among other parameters, the proportion of active and inactive baskets, borrowing limits, and number of baskets ranging from few to many.

Results

Operating characteristics are substantially affected by the number of baskets included in the analysis. Both limiting mechanisms mitigate these effects, though neither consistently outperforms the other across all simulated scenarios. These findings indicate that it may be appropriate to adjust the analysis when evaluating a large number of baskets.

Discussion

The number of baskets is an important, often overlooked design parameter in basket trials. ESS-based borrowing limits offer an interpretable and effective strategy to control the inappropriate influence of many baskets, while not being limited to the use of designs based on Power Priors. Investigators should carefully consider borrowing constraints during the design phase to maintain appropriate operating characteristics.

References:

Baumann, L., Sauer, L. D., & Kieser, M. (2024). A Basket Trial Design Based on Power Priors. Statistics in Biopharmaceutical Research, 1-11. https://doi.org/10.1080/19466315.2024.2402275

Chen, X., Zhang, J., Yang, W., Jiang, L., Li, B., & Yan, F. (2025). Enhancing interpretability for Bayesian basket trial designs by effective sample size. BMC Medical Research Methodology, 25(1), 278. https://doi.org/10.025-02715-x

---

Estimating Subtrial-specific Optimal Biological Doses in Basket Trials Permitting Robust Borrowing of Information

11:45-12:00

This is one of this year's ISCB StCA award-winning abstracts.

Presenter: Zhi Cao

Authors: Zhi Cao [1]; Pavel Mozgunov [1]; Haiyan Zheng [2]

Affiliations: [1] MRC Biostatistics Unit, University of Cambridge, Cambridge; [2] Department of Mathematical Sciences, University of Bath, Bath

Introduction

Drug development in the era of precision medicine increasingly uses basket trials and other multi-subgroup designs, where targeted therapies are evaluated across biomarker-defined patient subtrials. For targeted agents and immunotherapies, the objective in early development is no longer the maximum tolerated dose (MTD), but the optimal biological dose (OBD) that achieves the best

benefit-risk trade-off. Identifying an OBD separately within each subtrial is challenging because sample sizes in early-phase trials are small, and toxicity and efficacy need to be considered jointly.

Methods

We develop a Bayesian framework for subtrial-specific OBD selection that enables robust borrowing of information on two-dimensional parameters representing toxicity and efficacy. Building on the bivariate exchangeable-non-exchangeable (E-BiEXNEX) model [1, 2], we specify priors on subtrial-specific dose-toxicity and dose-efficacy parameters so that information sharing is data-adaptive and resistant under extreme observations. The prior distributions are calibrated across multiple candidates to promote robust performance over a wide range of plausible scenarios. Our primary setting assumes binary toxicity and continuous efficacy, but the approach can be extended to other endpoint types with minor modifications to the regression components. The dose recommendation is driven by a utility-based decision rule that combines posterior toxicity probabilities and efficacy means under pre-specified safety and futility constraints, yielding subtrial-specific OBD recommendations with additional important operating characteristics.

Results

Operating characteristics are evaluated under scenarios with varying degrees of between-subtrial similarity, dose-response shapes, and locations of true OBD compared to stand-alone (NEX only) analysis. The results show that information borrowing can improve the proportion of correct selection (PCS) on most scenarios and reduce the overly toxic selection, but occasionally over borrowing can happen and harm PCS.

Discussion

Enabling the information borrowing is generally advantageous compared with NEX only, but in the future we still need to pay close attention to erroneous borrowing patterns.

1Neuenschwander, B., Wandel, S., Roychoudhury, S., & Bailey, S. (2016). Robust exchangeability designs for early phase clinical trials with multiple strata. Pharmaceutical Statistics, 15(2), 123-134. https://doi.org/10.1002/pst.1730

2Cao, Z., Mozgunov, P., & Zheng, H. (2025). Robust Bayesian hierarchical models for basket trials enabling joint evaluation of toxicity and efficacy. https://arxiv.org/abs/2505.10317

Beyond the Odds Ratio: A Penalized Mantel-Haenszel Approach to Adaptive Borrowing in Basket Trials

12:00-12:15

Presenter: Xingzhi Xu

Authors: Xingzhi Xu; Satoshi Hattori

Affiliations: The University of Osaka, Suita

Introduction

Early-phase oncology basket trials evaluate targeted therapies across multiple independent tumor subtypes often in single-arm setting. Due to inherently small sample sizes, statistical methods that borrow information across similar subgroups are essential. Some papers successfully introduced the lasso technique to determine the amount of information to borrow. Current frequentist approaches rely on penalized logistic regression, with which baskets of similar response rate would borrow information from each other. A more natural clinical assumption, however, is that the therapy confers a consistent improvement across all baskets regardless of their baseline sensitivity, a scenario that current approaches cannot accommodate. In addition, these methods model the odds ratio, which suffers from non-collapsibility and is highly susceptible to computational instability when a cohort observes zero responders.

Methods

Instead of logistic regression likelihood, we propose a novel Penalized Mantel-Haenszel (MH) -type estimating equation accounting for heterogeneous baseline sensitivity over bakstes. By formulating a Weighted Least Squares objective combined with a fused group penalty, our method dynamically borrows information across subgroups based on the similarity of their observed treatment effects. Because it operates on the identity link, the proposed method evaluates a collapsible, clinically interpretable estimand and natively accommodates baseline heterogeneity across tumor types. To align with regulatory standards for complex adaptive designs, the penalization parameter and critical efficacy boundaries are pre-calibrated via simulation to guarantee strict, exact control of the Family-Wise Error Rate (FWER).

Results

Extensive simulation studies were conducted to compare the Penalized MH method against penalized Maximum Likelihood Estimation (MLE) approaches under both global null and sparse alternative scenarios. The Penalized MH method achieved comparable marginal power and Mean Squared Error to the exact MLE models even when baseline sensitivity is homogeneous. Crucially, the proposed method exhibited superior computational stability it demonstrated absolute immunity to complete separation, seamlessly computing bounded treatment effects in sparse cohorts where traditional logistic-based penalties failed to converge or required ad-hoc continuity corrections.

Discussion

The Penalized MH framework provides a highly stable and robust alternative for data-driven borrowing in basket trials. By directly penalizing the risk difference rather than the odds ratio, the method bypasses the boundary violations of small-sample binomial data, offering trialists a highly interpretable, regulatory-compliant methodology for early-phase oncology designs.

Time and Cost Effectiveness of Seamless Phase II/III Designs: A Simulation-Based Evaluation

12:15-12:30

Presenter: Vera Baumeister

Authors: Vera Baumeister; Alexander Pieper

Affiliations: Evidenze Germany GmbH, Essen

Introduction

Adaptive clinical trial designs allow prespecified interim adaptations such as early stopping for efficacy, for futility or sample size reassessments based on accumulating data. These procedures may shorten the operational duration of a study but require extensive statistical simulation, regulatory interaction, and cross-functional preparation, potentially prolonging the setup phase. Quantitative evidence describing how additional preparation time compares to potential time and cost reductions remains limited.

Methods

A modular R-based simulation framework was developed using established principles for adaptive design evaluation. The framework simulates patient level randomization and accrual times, applies one sided interim and final tests, models dose selection rules, and accounts for recruitment variability, early stopping, and operational costs. Both non-seamless and seamless Phase II/III adaptive development programs were represented using combined-evidence testing procedures and either continuous or staged operational timelines. Each scenario was evaluated using 10,000 Monte-Carlo simulations to obtain distributions of study duration, total cost, sample size, and success probabilities. The analysis quantifies time and cost implications under varying assumptions about the length of the adaptive setup phase.

Results

The full set of scenarios is still under evaluation. A representative analysis compared a traditional development program (Phase-II setup: 8 months transition to Phase III: 10 months) with seamless Phase II/III adaptive programs whose setup durations ranged from 10 to 18 months.

For the seamless design, 40 patients in Phase II and 130 patients in Phase III were evaluated, saving approximately 20 patients per arm compared with the traditional model. This configuration yields 90% Phase III power when the better-performing arm is selected and an overall success probability of 81.6% for the combined Phase II/III program. In this scenario, the adaptive design produced approximately 10 months of time savings from Phase-II planning start to Phase-III database lock. Full scenario results will be presented upon completion.

Discussion

Preliminary findings suggest that adaptive designs can meaningfully shorten overall development timelines and reduce operational costs, even when longer preparation phases are required. These results illustrate the potential strategic value of adaptive methodologies in clinical development.

Completion of the full simulation set will further clarify when adaptive approaches provide practical advantages.

Causal Estimands and Biases

Wed, 30.09.26, 11:00–12:30, Room: N1

Chair: Richard Post

Causal Inference Methods for Estimating Treatment Effects in the Presence of Rescue Medication and Treatment Discontinuation

11:00-11:15

Presenter: Louise Jespersen

Authors: Louise Jespersen [1]; Susanne Urach [2]; Paula Starke [3]; Tobias Fellinger [2]; Florian Klingmueller [2]; Franz Koenig [4]; Maddalena Centanni [5]; Tim Friede [6]; Angelika Geroldinger [2]; Andrew C. Hooker [5]; Maxi Schulz [7]; Robin Ristl [4]

Affiliations: [1] Medical University of Vienna, Center for Medical Data Science, Institute of Medical Statistics, Vienna; [2] Austrian Medicines and Medical Devices Agency (MEA / AGES), Vienna; [3] Institut für medizinische Statistik, Universitätsmedizin Göttingen, Göttingen; [4] Medical University of Vienna, Vienna; [5] Uppsala University, Uppsala, Sweden; [6] Universitätsmedizin Göttingen, Göttingen; [7] Universitätsmedizin Göttingen, Institut für Medizinische Statistik, Göttingen

Introduction

As highlighted in ICH E9(R1) Addendum, precise specification of the clinical question of interest is essential in randomised clinical trials. Preserving a causal interpretation of treatment effects requires careful alignment between the clinical objective, the strategy for handling intercurrent events(ICEs), and the statistical analysis. This work evaluates causal inference methods for handling ICEs in randomised clinical trials and compares their performance with commonly used analysis approaches.

Methods

We consider a diabetes trial setting in which treatment discontinuation and rescue medication constitute relevant ICEs. Both a treatment policy strategy and a hypothetical strategy are investigated.

The causal inference approaches inverse probability weighting (IPW), de-mediation, and the g-formula are compared with mixed models for repeated measures(MMRM) and multiple imputation(MI). IPW relies on pre-ICE outcome data and reweights individuals, who did not yet experience the ICE, to account for informative ICE occurrence. De-mediation treats ICEs as mediators and removes their contribution to the treatment effect.

The g-formula models the joint distribution of outcomes and ICE indicators to estimate treatment effects under specified interventions on both treatment and ICE.

A simulation study was conducted covering missing data mechanisms, sample sizes of 100-400 participants, and uneven distribution of rescue medication across treatment groups.

Results

The simulation study evaluates the performance of the methods in terms of bias, empirical standard error, and confidence interval coverage. Robustness to assumption violations and edge-case scenarios, such as violations of the positivity assumption due to perfect separation, is also assessed.

Discussion

This work provides methodological insight into the use of causal inference methods for handling ICEs within the estimand framework and forms part of a broader project investigating additional ICE scenarios and strategies.

The presentation has been produced in part by making use of the results under the specific contract no.03-FWC-EMA/2020/46/TDA/L3.02 managed by MU-Vienna, in its role as coordinator of the CONFIRMS-consortium. This presentation expresses the opinion of the authors and may not be understood or quoted as being made on behalf of or reflecting the position of the European Medicines Agency or one of its committees or working parties. This study has been registered in the HMA-EMA-Catalogue of Real-World-Data under the EUPAS1000000720.

Partial identification of causal effects in the presence of multiple versions of treatment

11:15-11:30

Presenter/Author: Chang Wei

Affiliations: Erasmus MC, Rotterdam

Introduction

To well define a causal question when there are multiple versions of treatment and estimate a causal effect, the treatment variation irrelevance assumption (TVIA) is required: all versions of the treatment are assumed to have the same effect on the outcome for all units. However, hidden issues arising from the violation of TVIA are common in practice, particularly in observational studies where detailed information on treatment is often unavailable. Examples include different types or dosages of medications, varying intensity or duration of physical activity, and various strategies for blood pressure control in pragmatic trials. Although a framework for treatment with multiple versions (compound treatment) has been proposed, it relies on a consistency assumption linking the treatment to its versions, and the target estimand represents an overall effect of the compound treatment. In contrast, in applied settings where only specific versions of a treatment are to be administered, the effect of a version-specific treatment is of primary interest.

Methods

When individual-level information on treatment versions is unavailable, in the context of either randomized or observational studies with a dichotomized treatment with multiple versions within one treatment arm, we derive large-sample bounds for version-specific treatment effects.

Results

We show that, for a given treatment version, the bounds can be obtained using knowledge of a lower bound on the proportion of that version in the study population, which may be obtained from external data or domain expertise. These bounds can be further narrowed under additional nonparametric assumptions.

Discussion

Finally, we illustrate the proposed method using an example on the effect of antidiabetic medications on cognitive decline, demonstrating how multiple dosages and types of antidiabetic treatments influence both estimation and interpretation of the results.

Adjusted Marginal Inference under Confounding

11:30-11:45

Presenter: Susanne Dandl

Authors: Susanne Dandl; Elise Dumas; Torsten Hothorn

Affiliations: Universität Zürich, Zurich

Introduction

A common approach to obtain unbiased causal effects in the presence of confounding is to include all confounders as covariates in a conditional outcome model. This covariate adjustment requires caution for noncollapsible effect measures - such as Cohen's d in linear regression, odds ratios in logistic regression, or hazard ratios in Cox models. Due to noncollapsibility, adding covariates to a model can change the interpretation of the estimated treatment effect, impeding comparisons between models with different adjustment sets.

Methods

Nonparanormal adjusted marginal inference (NAMI), introduced in [1], provides a general framework to address noncollapsibility. It estimates a marginal treatment effect whose interpretation is not influenced by covariates, while still benefiting from improved inference through their inclusion. Since it is based on transformation models, it can estimate marginal effects for diverse outcome types. However, because NAMI was originally developed for randomized data, it is unclear whether these marginal effect estimates remain unbiased under confounding.

Results

Our theoretical assessment is threefold. First, we show that NAMI can be understood as an extension of the conditional linear outcome model. Second, we demonstrate that NAMI recovers interventional distributions. Third, NAMI can be framed as a frugal parameterization of the joint distribution of outcome, treatment, and covariates introduced in [2]. We confirm these findings empirically in

simulation studies using diverse outcome types. NAMI estimates marginal effects under confounding with high precision and appropriate coverage compared to other adjustment methods. Clinical applications further demonstrate that NAMI enables the ranking of covariates by their prognostic strength and further insights into confounding mechanisms.

Discussion

Overall, this shows that NAMI provides unbiased estimators of practically relevant treatment effect measures in observational settings with confounding, particularly noncollapsible effect measures. Extensions to estimate heterogeneous treatment effects, as well as assessments of robustness under model misspecification, display important directions for future research.

[1] Dandl S., Hothorn, T. (2025). Nonparanormal adjusted marginal inference (with discussion). Biometrics (in press). Preprint: doi:10.48550/arXiv.2503.01657.

[2] Evans, R. J., Didelez, V. (2024). Parameterizing and simulating from causal models. Journal of the Royal Statistical Society. Series B, Statistical Methodology, 86(3), 535-568. doi:10.1093/jrsssb/qkad058.

Inverse Probability Weighting for Causal Inference in Hierarchical Data: A Comparison of Multilevel Propensity Score Strategies

11:45-12:00

This is one of this year's ISCB CASP award-winning abstracts.

Presenter: Lin Hu

Authors: Lin Hu; Chunxia Yang; Tao Zhang

Affiliations: Department of Epidemiology and Health Statistics, West China School of Public Health/West China Fourth Hospital, Sichuan University, Chengdu, China, Chengdu

Introduction

Estimating causal effects from hierarchical observational data is challenging when unmeasured cluster-level confounders are present. Inverse probability weighting (IPW) based on multilevel propensity score models is widely used, but its performance may be affected by model misspecification, covariate imbalance, and extreme weights. We compared alternative multilevel propensity score strategies for estimating average treatment effects (ATEs) in hierarchical data.

Methods

We conducted a simulation study under 48 hierarchical scenarios defined by different design types, cluster structures, positivity conditions, and treatment assignment mechanisms. Nine propensity score weighting strategies were evaluated, combining fixed-effects, random-effects, and Bayesian additive regression tree (BART) models with either marginal stabilized weights or cluster-mean stabilized weights. Extreme weights were additionally handled by truncation. Performance was assessed using the absolute standardized mean difference, bias, absolute percentage bias, cluster-robust

standard error, empirical standard error, root mean square error (RMSE), 95% confidence interval coverage, and weight distribution. The methods were also applied to data from patients co-infected with HIV and tuberculosis (TB) in Liangshan Prefecture, China, to evaluate the causal effect of TB treatment delay on treatment outcomes.

Results

The simulation study showed that FEM-Marginal tended to generate the most extreme weights, whereas BART-FE-Marginal substantially reduced extreme weights in a large number of small clusters. When the data satisfied the positivity assumption, the marginal stabilized weight strategy showed the largest absolute percentage bias and RMSE, whereas the cluster-mean stabilized weight strategy showed the smallest. In the case study, different ATE strategies consistently showed that TB treatment delay was a risk factor for treatment outcome among HIV-TB co-infected patients.

Discussion

For hierarchical data with unmeasured cluster-level confounding, incorporating cluster characteristics into propensity score estimation is critical for improving causal effect estimation. BART-based multilevel propensity score models and cluster-mean stabilized weights appear promising, whereas additional handling of extreme weights remains important when marginal stabilized weights are used. In hierarchical data with unmeasured cluster-level confounders, reducing extreme weights, weight variability, and model misspecification while improving covariate balance can effectively minimize estimation bias.

Selection bias due to causal effect heterogeneity in instrumental variable analyses

12:00-12:15

Presenter: Apostolos Gkatzionis

Authors: Apostolos Gkatzionis [1]; Fernando Hartwig [2, 1]; Kate Tilling [1]

Affiliations: [1] University of Bristol, Bristol; [2] Federal University of Pelotas, Pelotas

Introduction

Selection bias is a common concern in causal inference. Two mechanisms can give rise to selection bias: stratifying on a collider of the exposure and outcome, and stratifying on an effect modifier of the exposure-outcome causal effect. Instrumental variable (IV) analysis is an approach for causal inference that can mitigate bias due to confounding and reverse causation. However, IV analyses remain susceptible to selection bias. Research into the impact of selection bias on IV analyses has focused primarily on collider bias.

Methods

We show that IV estimates of the average causal effect (ACE) are biased if selection (or a cause of selection) is an effect modifier for the ACE. This can happen even if the "no simultaneous heterogeneity" assumption, which is required to identify the ACE in IV studies, is satisfied. In a simulation study, we investigate the size and direction of the bias when estimating effects on the mean difference or odds ratio scale, assuming effect modification by either a confounder of the IV exposure and outcome or an unrelated variable. We also investigate bias due to heterogeneity in the instrument-exposure relationship. Finally, we explore diagnostics and methods to deal with the bias in practice.

Results

Our simulations illustrate that effect modification of the exposure-outcome causal effect can induce selection bias in IV analyses. The magnitude of bias depends on the strength of effect modification and the strength of association of the effect modifier with selection. Bias is not induced under the (sharp) null hypothesis of no exposure-outcome causal effect. When bias occurs, inverse probability weighting can be used to mitigate it, provided that the effect modifier is observed. On the other hand, heterogeneity in the instrument-exposure association does not affect the standard two-stage least squares IV estimator due to its robustness to misspecification of the first-stage model.

Discussion

Selection bias can be induced in IV analyses if an effect modifier of the exposure-outcome effect also affects selection. Researchers working on IV analyses should be aware of this type of bias and take appropriate steps to account for it.

Beyond the trees: Explainable Bayesian Causal Inference with semi-parametric BART

12:15-12:30

Presenter: Yusufhan Balci

Authors: Yusufhan Balci [1]; Mark van de Wiel [2]; Jeroen Hoogland [1]

Affiliations: [1] Amsterdam UMC, Amsterdam; [2] Amsterdam University medical centers, Amsterdam

Introduction

Estimating and inferring subgroup-specific treatment effects, or the Conditional Average Treatment Effect (CATE), is a classic challenge in personalized medicine. While contemporary causal machine learning methods can flexibly predict the CATE, they often act as 'black boxes,' offering limited interpretability and struggling to provide rigorous inference on which patient characteristics actually drive treatment heterogeneity. We argue that accurate, interpretable predictions of these effects are essential for transparent medical decision-making.

Methods

To combine the flexible prognostic estimation of machine learning with interpretable inference, we propose a novel semi-parametric Bayesian framework. Our approach leverages a non-parametric Bayesian Additive Regression Trees (BART) prior to capture complex, non-linear baseline risks, rendering the prognostic function robust to misspecification. Crucially, we isolate the treatment heterogeneity using a linear model with interactions, endowed with a sparse Horseshoe prior. This structural separation allows for direct parameter inference and explicit variable selection. Finally, we employ Shapley values to computationally quantify the substantive additive contribution of each covariate and its interactions to the final CATE estimate.

Results

Simulation studies demonstrate that our semi-parametric approach yields significant efficiency gains and vastly narrower credible intervals when the true treatment effect has a linear or homogeneous structure, successfully extracting signal where fully non-parametric models struggle at limited sample sizes. When applied it to a clinical trial program, our method matched the out-of-sample predictive accuracy and clinical ranking power of fully flexible models. More importantly, using Shapley value decomposition, our model effectively acted as a strict regularizer it cleanly isolated covariates as active effect modifiers while rigorously shrinking the contributions of non-informative baseline variables to zero, preventing the identification of spurious subgroups.

Discussion

The proposed semi-parametric framework successfully bridges the gap between advanced causal machine learning and clinical utility. By explicitly separating the prognostic and treatment components, our method achieves high predictive accuracy while uniquely providing transparent, individualized insights into why a specific patient is expected to benefit from a targeted intervention.

Federated Health Data Systems

Wed, 30.09.26, 11:00–12:30, Room: N2

Chair: Jessica Vasseur

A Locally Deployable FHIR Validation Service for Federated Health Data Infrastructures - A Technical Case Report

11:00-11:15

Presenter: Julian Saß

Authors: Julian Saß [1]; Alexander Kiel [2]; Joshua Wiedekopf [3]; Julian Gruendner [4]; Maximilian Kurscheidt [5]

Affiliations: [1] Berliner Institut für Gesundheitsforschung in der Charité - Universitätsmedizin Berlin, Berlin; [2] Universität Leipzig, Leipzig; [3] Universität zu Lübeck, Lübeck; [4] Lehrstuhl für Medizinische Informatik, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen; [5] Hochschule Heilbronn, GECKO Institut für Medizin, Informatik und Ökonomie, Heilbronn

Introduction

The Medical Informatics Initiative enables the secondary use of clinical routine data in German university medicine through a federated FHIR-based infrastructure, with implementation guides defining the underlying data representation. Reliable validation of FHIR resources is required to ensure conformance with these specifications. However, validation in federated environments remains challenging due to heterogeneous infrastructures, limited terminology availability, and difficulties in establishing reproducible validation setups.

Methods

This technical case report describes the design and implementation of a locally deployable FHIR validation service in the MII context. The system combines the HL7-maintained FHIR Validator with a configurable terminology service and a containerized deployment model. Structural and terminology validation are separated by design. Implementation guides and terminology resources are provisioned in a reproducible manner. The system supports both local terminology deployment and integration with existing institutional terminology services.

Results

The proposed solution enables automated conformance testing of FHIR resource instances against MII implementation guides in local and federated environments. The service is provided as a reusable containerized validation component that can be integrated into development and data processing workflows, improving reproducibility, and reducing configuration effort across sites.

Discussion

The results demonstrate that reproducible FHIR validation can be achieved through standardized deployment and controlled terminology provisioning. Remaining limitations are primarily related to incomplete or unavailable terminology content. Further work is required to evaluate the approach in production settings across data integration centers and to assess integration into data processing pipelines, including the handling and interpretation of validation results.

The full paper of this abstract will be published in the GMDS series "Studies in Health Technology and Informatics" with doi: 10.3233/SHTI260987, URL:https://dx.doi.org/10.3233/SHTI260987

FHIRJoin: A Prototypical Concept for Federated Queries Across Distributed FHIR Servers

11:15-11:30

Presenter: Marvin Schmidt

Authors: Marvin Schmidt [1]; Abishaa Vengadeswaran [1, 2]; Michael Folz [1]; Sara Bachir [2]; Andreas Heidenreich [2]; Dennis Kadioglu [1, 2]; Holger Storf [1, 2]

Affiliations: [1] Goethe University Frankfurt, University Medicine Frankfurt, Institute of Medical Informatics (IMI), Frankfurt am Main; [2] University Medicine Frankfurt, Data Integration Center (DIC), Frankfurt am Main

Introduction

The Health Level 7 Fast Healthcare Interoperability Resources (FHIR) standard has become the widely adopted framework of the German Data Integration Centers (DIC) that enables interoperable access to healthcare data through standardized REST-based interfaces. In many real-world research environments, however, clinical data are often distributed across multiple independent FHIR servers. This limits the ability to execute automated search queries that span multiple data repositories. Standard FHIR servers lack native capabilities to perform aggregate distributed searches resulting in significant hurdles for researchers in cohort discovery.

Methods

This paper presents FHIRJoin, a concept and prototype for enabling federated FHIR queries across distributed FHIR servers. A Python-based prototype was developed using the FastAPI framework which decomposes complex search queries into server-specific subqueries. It retrieves patient identifiers from each server and aggregates the results using set-based operations.

Results

The resulting patient cohort is ultimately used to retrieve and return a FHIR resource bundle to the client. FHIRJoin was evaluated in a distributed setup with multiple FHIR servers containing synthetic clinical data by utilizing queries provided by the DIC Frankfurt.

Discussion

The proposed approach provides a scalable and lightweight solution for federated cohort discovery without centralized data integration, making it suitable for distributed healthcare research infrastructures.

The full paper of this abstract will be published in the GMDS series "Studies in Health Technology and Informatics" with doi: 10.3233/SHTI260984, URL:https://dx.doi.org/10.3233/SHTI260984

Interoperable Clinical Data Integration for Distributed Medical AI: The Open Medical Inference Gateway

11:30-11:45

Presenter: Kaushik Manjunatha, Nimesh Singh

Authors: Obioma Pelka [1, 2]; Jan Eil [1]; Kaushik Manjunatha [1]; Nimesh Singh [1]; Jazia Omeirat [2]; Stefan Sigle [3]; Georg Mathes [4]; Simon Tobias Schweizer [5]; Shura Roman Stump [6]; Gerda Girdziunaite [7]; Felix Nensa [8]

Affiliations: [1] Institute for Artificial Intelligence in Medicine (IKIM), University Hospital Essen, Essen; [2] Central IT Department, Data Integration Center, University Hospital Essen, Essen; [3] MOLIT Institute gGmbH, Heilbronn; [4] MOLIT Institut gGmbH, Heilbronn; [5] GECKO Institute for Medicine, Informatics and Economics, Heilbronn University of Applied Sciences, Heilbronn; [6] Department of Radiology, University Medical Center Freiburg, Freiburg; [7] Charité - Universitätsmedizin Berlin, Berlin; [8] Institut für künstliche Intelligenz in der Medizin (IKIM), Universitätsklinikum Essen, Universität Duisburg-Essen, Essen

Background: Integrating AI inference services into clinical workflows remains a persistent challenge, primarily because inference backends and healthcare data standards operate with fundamentally different assumptions. The EU AI Act further requires transparency and traceability for high-risk AI systems in healthcare. Methods: This paper presents the Open Medical Inference (OMI) Gateway, an interoperability layer that mediates between FHIR- and DICOM-based clinical data environments and heterogeneous AI inference backends. The gateway was implemented in Python using FastAPI with asynchronous job processing via Celery and RabbitMQ. The architecture follows a service-oriented plugin model that separates shared interoperability concerns from service-specific transformation logic. Input and output data are validated against FHIR profiles and DICOM specifications at multiple pipeline stages, producing OperationOutcome resources as a machine-readable audit trail. Results: End-to-end pipeline latency was measured across 100 sequential inference requests, yielding a median latency of 263 ms and a 100% success rate. The design was evaluated against interoperability requirements from German Medical Informatics Initiative (MII) Data Integration Centers. The system is deployable with a single command and available as open-source software. Discussion: The results indicate that embedding validation and structured artifact handling directly into the inference execution path yields a more governable and auditable integration model than ad-hoc service adapters.

The full paper pertaining to this abstract will be published in the GMDS journal "Medizinische Informatik, Biometrie und Epidemiologie (MIBE)."

An Interoperable Platform for AI-Based Analytics of Wearable Data

11:45-12:00

Presenter: Elias Grünewald

Authors: Elias Grünewald; Louis Agha-Mir-Salim; Felix Balzer

Affiliations: Charité - Universitätsmedizin Berlin, Berlin

Introduction

Wearable devices enable continuous capture of physiological parameters in real-world settings and offer potential for personalized medicine, prevention, and in-hospital analytics. However, key challenges remain, including heterogeneous data landscapes, limited interoperability, and insufficient integration into clinical information systems.

Methods

We designed and implemented a microservice-based platform architecture to standardize data integration from consumer and medical-grade wearable devices. The platform is event-driven and built on open standards, particularly FHIR. A declarative mapping layer enables dynamic transformation of wearable data into FHIR resources. Data from heterogeneous devices are harmonized through a semantic layer for ad hoc and longitudinal analytics. Clinical evaluation across multiple studies at Charité - Universitätsmedizin Berlin is currently in preparation.

Results

The platform enabled data integration from multiple wearable devices, including activity, heart rate, and sleep metrics. Data harmonization supported standardized analyses across vendor ecosystems, such as Apple Health and Google Fit. Initial evaluations demonstrated high data completeness, stable system availability, and scalability for thousands of concurrent devices. We also demonstrated integration with clinical information systems to combine wearable data with electronic health records. Compared with related approaches, the platform provides a unique combination of cloud-native stream processing and lakehouse analytics capabilities. In addition, it supports online machine learning pipelines for near-real-time use cases. Pilot results indicate fast and reliable insight generation, for example for data acquisition and interpretation during in-hospital six-minute walk tests. Furthermore, we show that the platform supports studies across in-hospital (e.g., in intensive care units), and ambulatory settings (e.g., in psychiatry) on the same analytics infrastructure. Our benchmarking identified message size as the major factor affecting end-to-end processing time, particularly as message sizes increase in asynchronous long-term observation scenarios.

Discussion

The results demonstrate the feasibility of an interoperable wearable platform for research and clinical practice. Open standards and modular architectures are key enablers. Limitations include device dependency and unresolved questions regarding regulatory classification. Overall, the platform provides a scalable foundation for integrating wearable data into medical information systems and can contribute to more effective data-driven care concepts. The platform will be released as open-source software.

A Multi-site Dashboard for Evaluating HiGHmed Cardiovascular Sensor Data from Apple Watch

12:00-12:15

Presenter: Sabine Hanß

Authors: Olusola Khalid Yusuf [1]; Tabea Steinbrinker [2]; Sabine Hanß [1]; Katharina Jörß [1]; Jendrik Richter [3]; Tanja Zeppernick [4]; Kim Katrin Werle [5]; Erik Tute [6]; Robert Freund [7]; Georg Fette [8]; Udo Bavendiek [4]; Dirk Grosse-Meininghaus [7]; Stefan Störk [9]; Dagmar Krefting [1]

Affiliations: [1] Department of Medical Informatics, University Medical Center Göttingen (UMG), Göttingen; [2] Department of Medical Informatics, University Medical Center Göttingen (UMG), 37075; [3] University Medical Center Göttingen, Department of Medical Informatics, Göttingen, Göttingen; [4] Department of Cardiology and Angiology, Hannover Medical School, Hannover; [5] Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Hannover; [6] Peter L. Reichertz Institut für Medizinische Informatik der Technischen Universität Braunschweig und der Medizinischen Hochschule Hannover, Hannover; [7] Medical University Lausitz - Carl Thiem, Cottbus; [8] Universitätsklinikum Würzburg, Würzburg; [9] Department Clinical Research & Epidemiology, Comprehensive Heart Failure Center Würzburg and Department of Internal Medicine I, University Hospital Würzburg, Würzburg

Introduction

The HiGHmed use case cardiology aims to enable the early detection of clinical worsening of patients with chronic heart failure. Two complementary studies are introduced, comprising: 1) a routine clinical study capturing in-hospital patient data, and 2) a sensor sub-study monitoring patients remotely through Apple Watches. Wearable devices continuously collect cardiovascular metrics, including heart rate and step count, and transmit data to participating medical data integration centers. This study presents a dashboard tool designed to support cardiologists in evaluating multi-site sensor data collaboratively.

Methods

We developed a centralized R Shiny dashboard that aggregates Apple Watch sensor data with hospitalization records from 12 of 93 participants across four study sites. The dashboard integrates device-recorded metrics (heart rate and steps), manually recorded clinical parameters (resting heart

rate, blood pressure, and weight), Kansas City Cardiomyopathy Questionnaire (KCCQ-12) scores, and hospitalization events. Three analytical modules were constructed: 1) an observation overview displaying stacked multi-metric trends, 2) a heart rhythm analyzer presenting hourly heart rate patterns with three-day rolling windows, and 3) a correlation module visualizing relationships between metrics using 14-day sliding windows.

Results

The observation overview displays pre-hospitalization deterioration, including low KCCQ-12 scores (e.g., ~50) and reduced activity levels (daily steps dropping below 2,000). Hourly heart rhythm analysis demonstrates a decline observed on hospitalization day compared to baseline. Correlation analysis confirmed expected physiological relationships, including a strong positive correlation between steps and device heart rate ($r \approx 0.8$). There are visible gaps in sensor recordings and inconsistent adherence to manual measurement, suggesting data quality challenges.

Discussion

The sensor data analysis dashboard provides a practical tool for multi-site cardiovascular data evaluation, enabling clinicians to transition from case-by-case parameter review to systematic multi-metric assessment of cardiac patients.

The full paper of this abstract will be published in the GMDS series "Studies in Health Technology and Informatics" with doi: 10.3233/SHTI260986, URL:https://dx.doi.org/10.3233/SHTI260986

Personalized PREdiction of Transplant ToxicitY through federated learning from data, expert opinions and patient perspectives (PRETTY)

12:15-12:30

Presenter: Yasmine Alwie

Authors: Yasmine Alwie [1, 2]; Danny Ammon [3]; Prateek Anand [4, 5]; Alexander Casimir Angleitner [6]; Maximilian Fleischmann [7]; Georg-Nikolaus Franke [8]; Han Fu [9]; Michael Heuser [10]; Inken Hilgendorf [7]; Franziska Jahn [10]; Tim Kacprowski [4, 5]; Tibor Kesztyues [11]; Lutz P. Müller [10]; Emetis Niazmand [4, 5]; Steffen Oeltze-Jafra [1]; Judith Schaffrath [10]; Daniel Tiller [12]; Thomas Wendt [13]; Gerald Wulf [6]; Guoren Xi [3]

Affiliations: [1] Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Hannover Medical School, Hannover, Germany., Hannover; [2] All authors are listed in alphabetical order., N/A; [3] Data Integration Center, Jena University Hospital, Jena, Germany, Jena; [4] Institute of Data Science in Biomedicine, Technische Universität Braunschweig, Braunschweig, Germany, Braunschweig; [5] Braunschweig Integrated Centre of Systems Biology (BRICS), TU Braunschweig, Braunschweig, Germany, Braunschweig; [6] Department of Hematology and Medical Oncology, University Hospital Göttingen, Göttingen, Germany, Göttingen; [7] Klinik für Innere Medizin 2, Abteilung für Hämatologie und Internistische Onkologie, Universitätsklinikum Jena,

Jena, Germany, Jena; [8] Department of Hematology, Hemostaseology and Cellular Therapy, University of Leipzig Medical Center, Leipzig, Germany, Leipzig; [9] Data Integration Center, University Hospital Halle (Saale), Halle, Germany, Halle (Saale); [10] Department of Internal Medicine IV, University Hospital Halle (Saale), Martin-Luther-University Halle-Wittenberg, Halle, Germany, Halle (Saale); [11] Medical Data Integration Center, Department of Medical Informatics, University Medical Center Göttingen, Göttingen, Germany, Göttingen; [12] Data Integration Center, University Hospital Halle (Saale), Halle (Saale); [13] Center for Medical Informatics, Data Integration Center, University of Leipzig Medical Center, Leipzig, Germany, Leipzig

Introduction

Authors listed in alphabetical order. Allogeneic hematopoietic cell transplantation (allo-HCT) is curative for many leukemia patients but frequently complicated by severe nephrotoxicity. Accurate, patient-specific risk prediction could guide preventive strategies, yet developing such models requires large, interoperable datasets and is hampered by data heterogeneity, fragmentation across institutions, and strict data-protection regulations. The PRETTY consortium consists of six German universities (Hannover, Braunschweig, Halle, Leipzig, Göttingen, Jena). It addresses these challenges by making data interoperable between sites, using federated learning of Bayesian networks (BNs), and integrating expert knowledge and patient perspectives.

Methods

Clinical variables are retrospectively extracted from the EBMT registry and from site-specific laboratory/nephrotoxicity databases for integration into the local Data Integration Center (DIC). Patient-reported outcomes (PROs) are collected prospectively through validated quality-of-life questionnaires. Data are discretized and integrated into the DIC as interoperable HL7 FHIR resources. Local discrete BNs are trained per site and then combined into a global model through federated learning. Clinical experts contribute prior knowledge to initialize network structure and perform iterative validation.

Results

We retrospectively included 3011 adult patients with acute leukemia or myelodysplastic syndrome who received allo-HCT between 01-01-2010 and 31-03-2024 in any of the participating clinics. We processed EBMT-datasets (210 features) and evaluated availability of laboratory/nephrotoxicity-dataset variables (131 features) across DICs, which revealed limited availability. Patient representatives actively shaped the PRO data collection scheme (53 features). FHIR profiles for all variables were defined, data processing and integration are ongoing. To facilitate model learning, we implemented a novel federated framework for discrete BN learning using FeatureCloud. Data of the prospective cohort ($n \approx 100$, recruitment ongoing) will be used for external validation.

Discussion

Our federated framework provides a privacy-preserving solution to use fragmented clinical data. It contributes to the generation of transparent risk predictions by exposing BN structure, which allows to trace predictions back to risk factors. Integration of physicians' and patients' perspectives will enhance reliability. Funding: German Federal Ministry of Research, Technology and Space (Grants

01KD2416A-F) Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 527049502.

Diagnostic Estimands and Modeling

Wed, 30.09.26, 11:00–12:30, Room: N4

Chair: Naoki Ishizuka

Application of the estimand concept in diagnostic accuracy studies

11:00-11:15

Presenter: Antonia Zapf

Authors: Daniel Dümmler [1]; Alexander Fierenz [2]; Philipp Weber [2]; Frederike Vogel [2]; Nicole Rübsamen [1]; Antonia Zapf [2]

Affiliations: [1] Institute of Epidemiology and Social Medicine, University of Münster, Münster; [2] Department of Medical Biometry, University Medical Center Hamburg-Eppendorf, Germany, Hamburg

Introduction

In 2020, the Estimand Framework was introduced for therapeutic trials in the appendix to the ICH E9 guideline [1]. The estimand is used to precisely define the treatment effect and, amongst other things, to define strategies for events that influence the interpretation or existence of measurements. In 2025, Fierenz et al. presented a counterpart Estimand Framework for diagnostic accuracy studies [2]. This presentation aims to illustrate the concept and its necessity using an example study.

Methods

For the study by Luetkens et al. (2025) [3], different estimands were defined based on the attributes of population, target condition, index test, accuracy measure and strategies for dealing with interfering events. The data were then analysed according to these estimands with sensitivity and specificity as co-primary endpoints and the results were compared.

Results

The different estimands reflect different study objectives and research questions. Accordingly, sensitivity varies greatly between 57% (confidence interval CI: 26% - 88%) and 88% (CI: 80% - 98%), whilst specificity remains almost unchanged between 95% (CI: 91% - 99%) and 97% (CI: 94% - 100%).

Discussion

This example application illustrates how important it is, when planning a diagnostic accuracy study, to carefully consider what the study objective actually is, to define the estimand accordingly, and also to consider what events may occur and how they should be handled.

References

[1] EMA (2020). https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e9-r1-addendum-estimands-and-sensitivity-analysis-clinical-trials-guideline-statistical-principles-clinical-trials-step-5_en.pdf

[2] Fierenz A et al. The Estimand Framework in Diagnostic Accuracy Studies. Stat Med. 2025 Sep44(20-22):e70248. doi: 10.1002/sim.70248.

[3] Luetkens KS et al. Gantry-free cone-beam CT arthrography for diagnosis of scapholunate ligament injuries: accelerating the preoperative work-up in acute wrist trauma. Eur Radiol. 2025 Aug35(8):4562-4570. doi: 10.1007/s00330-025-11405-7.

The joint Framework for dealing with missing values in combination with the estimand in diagnostic accuracy studies

11:15-11:30

Presenter: Alexander Fierenz

Authors: Alexander Fierenz; Katharina Stahlmann; Mahnaz Badpa; Antonia Zapf

Affiliations: University Medical Center Hamburg-Eppendorf, Hamburg

Introduction

Diagnostic accuracy studies aim to evaluate how precisely a developed test (called index test) can detect a certain health condition in comparison with a reference standard. During the study conduct, several issues may harm the validity of the result. Ignoring such interfering events could affect the accuracy estimation of the index test (Fierenz et al., 2025) or excluding missing values in the index test could lead to a misleading study result (Whiting et al., 2013). The estimand framework suggests to choose between different strategies based on the study objective (Fierenz et al., 2025) to handle interfering events. Missing values can be addressed by appropriate methods (Stahlmann et al., 2025). The simulation study aims to compare different estimands with varying methods for missing values to assess the sensitivity and specificity of a binary index test.

Methods

Results of the index test, reference standard and three covariates were simulated. Then, index test results were altered according to two generated interfering events and different missingness mechanisms. Three estimands were compared regarding relative bias and coverage probability. The naïve estimand reflects the standard approach, were only existing index test results are used and missing values are excluded from the analysis. The worst-case estimand evaluates interfering events and missing values as wrong test decision and represents the worst possible case. The advanced estimand defines suitable strategies for interfering events based on their characteristics and handles missing values with multiple imputation by chained equations.

Results

The naïve estimand and the advanced estimand show little bias under the missing completely at random (MCAR) assumption for sensitivity. Regarding specificity and coverage probability, the naïve estimand slightly outperforms the advanced estimand. Under missing at random (MAR), the advanced estimand outperforms the naïve estimand with respect to bias. All estimands deteriorate under missing not at random (MNAR). The worst-case estimand generally shows a high bias and low coverage probability.

Discussion

Overall, the advanced approach is the best way to get a precise estimation of study objective. The choice of the estimands beforehand is a complex but valuable process. Advanced methods should also be used to handle missing values.

Sample Size Determination for Optimal and Sub-Optimal Designs in Simplified Parametric Test Norming

11:30-11:45

Presenter: Francesco Innocenti

Authors: Francesco Innocenti [1]; Alberto Cassese [2]

Affiliations: [1] Maastricht University, Maastricht; [2] The University of Florence, Firenze

Introduction

Norm-referenced neuropsychological tests are widely used in clinical practice and diagnostic settings to detect cognitive impairment and related disorders. A prominent example is the Trail Making Test, which assesses multiple domains of cognitive functioning and has been shown to predict physical impairment and mortality in older adults. Clinical decisions often rely on cut-offs derived from normative models, yet instability in these norms due to sampling variability can lead to misclassification and reduced diagnostic reliability. Despite their critical role in high-stakes decision-making (e.g., the assignment to a treatment), formal guidance on sample size determination and optimal sampling design for parametric norming remains limited.

Methods

This work provides formulas, applicable to any sample composition, for determining the required sample size for univariate and multivariate normative studies under the simplified parametric norming framework. We further derive optimal sampling designs that minimize sampling error in estimated norms under (multivariate) multiple linear regression models.

Results

The derived closed-form formulas enable researchers to determine the required sample size under two criteria: (1) to ensure sufficient statistical power for classifying a patient's performance on a neuropsychological test, and (2) to estimate the classification cut-off within a prespecified precision level with a given assurance probability, following the Accuracy in Parameter Estimation (AIPE) approach. Optimal sampling designs were derived from 45 multivariate multiple linear regression models, assuming normality and homoscedasticity. These models varied in interaction structure among three common norm-predictors: a continuous variable (e.g., age), a categorical variable (e.g., sex), and a variable (e.g., education) treated as either continuous or categorical. To support practical implementation, three interactive R Shiny applications were developed. Their use is illustrated through the hypothetical planning of a normative study for the Trail Making Test, along with a review of the most commonly used models for this test in current practice.

Discussion

This work makes several contributions aimed at enhancing diagnostic accuracy in clinical practice by improving the design of studies used to derive reference values for neuropsychological tests.

Innocenti F., & Cassese, A. (2025). Sample Size Determination for Optimal and Sub-Optimal Designs in Simplified Parametric Test Norming. Multivariate Behavioral Research. DOI: https://doi.org/10.1080/00273171.

Accuracy and fairness comparison of raw and demographically-corrected scores

11:45-12:00

Presenter/Author: Ignacio Gonzalez Perez

Affiliations: EPFL, Lausanne

Introduction

In many applications, individuals are classified using numeric scores. For example, psychologists and physicians use screening tests to classify individuals as having cognitive impairment. Such scores are often "corrected" for fixed covariates, such as age, education and other demographics, before applying a diagnostic threshold. While such corrections are widespread, their motivation is usually "intuitive" rather than formal, and there is an ongoing debate about whether these corrections are desirable [1, 2].

Methods

We develop formal theory and methodology to study when demographic corrections are beneficial. Using tools from causal inference, yet without relying on causal models, parametric assumptions, or favourable empirical results, we show that raw scores can outperform corrected scores. Our analysis yields interpretable sufficient conditions under which raw scores dominate commonly used corrections. Furthermore, we study whether these corrections are compatible with formal notions of fairness.

Results

We applied our methods to the OASIS-3 dataset [3] to assess the performance of age-corrections of the Mini-Mental State Examination (MMSE) as a screening tool for cognitive impairment. Our results point towards the raw MMSE score having higher classification accuracy. Neither the raw nor the age-corrected MMSE scores achieve our targeted notion of fairness. Nevertheless, the age-corrected score is fair amongst the non-impaired. While this could have been expected due to the nature of the age-correction, it is still not enough to deem the corrected score fair.

Discussion

Our methods give formal results clarifying a long-standing debate about demographic corrections in psychological and medical testing. We debunk existing intuitions about the nature of demographic corrections and give an implementable method to assess whether they are appropriate.

[1] Piccininni, Marco, et al. 'Should cognitive screening tests be corrected for age and education? Insights from a causal perspective.' American journal of epidemiology 192.1 (2023): 93-101.

[2] Ilardi, Ciro Rosario, et al. 'No matter how big it is, but how you use it: the importance of demographic adjustment in clinical neuropsychology.' Neurological Sciences 46.2 (2025): 1027-1030.

[3] LaMontagne, Pamela J., et al. 'OASIS-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease.' medrxiv (2019): 2019-12.

Improved Polygenic Risk Modeling in Parkinson's Disease through Comparative Evaluation of Statistical Methods

12:00-12:15

Presenter: Sebastian Sendel

Authors: Sebastian Sendel [1, 2]; Zied Landoulsi [3]; Katja Lohmann [4]; Björn-Hergen Laabs [5, 4]; Inke R. König [6]; Patrick May [3]; Christine Klein [4]; Amke Caliebe [1, 2]

Affiliations: [1] Health and Medical University Erfurt, Erfurt; [2] Christian-Albrechts-University of Kiel, University Hospital Schleswig-Holstein, Campus Kiel, Kiel; [3] University of Luxembourg, Esch-sur-Alzette; [4] University of Luebeck, University Medical Center Schleswig-Holstein, Campus Luebeck, Lübeck; [5] University Medical Center Goettingen, Göttingen; [6] Universität zu Lübeck, Lübeck

Introduction

The etiology of Parkinson's disease (PD) is multifactorial, involving both genetic and environmental components. Polygenic scores (PGS) have been proposed to quantify genetic susceptibility. While conventional approaches such as clumping and thresholding are widely used, newer methods based on penalized regression and Bayesian inference may better account for the polygenic architecture of PD.

Methods

We analyzed genotype data from 1,762 idiopathic PD patients and 4,227 controls of European ancestry from the ProtectMove cohort (www.protectmove.de). Five PGS tools representing three statistical approaches were compared: clumping and thresholding (PRSice-2), penalized regression (lassosum2, LDAK), and Bayesian methods (LDpred2, PRS-CSx). Model performance was assessed using area under the curve (AUC). The best-performing PGS was externally validated in two independent datasets and further evaluated in 771 carriers of PD-associated variants (335 cases, 436 controls).

Results

Bayesian approaches showed the highest predictive performance, followed by penalized regression and clumping and thresholding. The best-performing model (LDpred2 928,814 SNPs) achieved an AUC of 0.680 (95% confidence interval: 0.665-0.695) in the ProtectMove cohort and 0.718 and 0.667 in external validation datasets, outperforming previously published scores. Among carriers of PD-related variants, the PGS yielded an AUC of 0.639 for GBA1 and 0.594 for heterozygous PRKN variant carriers.

Discussion

Bayesian methods appear particularly suitable for constructing PGS in PD. The developed PGS demonstrates improved predictive performance and may contribute to individualized risk stratification in idiopathic PD. Moreover, the findings suggest that polygenic risk influences disease manifestation even in individuals with known pathogenic variants, highlighting shared genetic mechanisms between idiopathic and monogenic forms of PD.

Developing a Machine Learning Framework for Focal vs. Non-Focal ARDS Phenotyping Using Chest X-Rays: A Multi-Cohort Validation Study

12:15-12:30

Presenter: Laurent Renard Triché

Authors: Laurent Renard Triché [1, 2, 3, 4]; Matthieu Jabaudon [1, 2]; Sylvie Chevret [5, 3]; Dominic C Marshall [4, 6]

Affiliations: [1] Department of Perioperative Medicine, CHU Clermont-Ferrand, Clermont-Ferrand; [2] iGReD, INSERM, CNRS, Université Clermont Auvergne, Clermont-Ferrand; [3] ECSTRRA Team, IRSL, INSERM UMR1342, Université Paris Cité, Paris; [4] Department of Surgery and Cancer, Imperial College London, London; [5] Department of Biostatistics, Hôpital Saint-Louis, AP-HP, Paris; [6] Cleveland Clinic London, London

Introduction

Acute Respiratory Distress Syndrome (ARDS), a lung condition, affects around 20% of intensive care units patients, with a high mortality. Patient heterogeneity hinders treatment discovery phenotyping may enable personalised therapy. The LIVE study (NCT02149589) investigated CT-defined

"focal"/"non-focal" phenotypes to guide ventilation however, approximately 20% of cases were misclassified.

While computed tomography (CT) remains the gold standard for radiological phenotyping, its utilisation is limited by cost, availability and transport risks for critically ill patient. Although Chest X-ray (CXR) is often the only practical alternative, existing CXR algorithms are not designed to differentiate these specific phenotypes.

Our primary objective was to develop a CXR-based model for ARDS classification.

Methods

Two cohorts were used: the LIVE study for model development, and an external cohort (Imperial College London, 3 UK hospitals) for validation.

Initially, CXRs were pre-processed, including cropping and normalisation, and data augmented due to the limited sample size. Several types of models were explored, including contrastive, pretrained and foundation models (with frozen or partially fine-tuned backbones), paired with simple classifiers such as Random Forest or XGBoost. Models were evaluated using cross-validation and three different random seeds.

The best-performing model was subsequently tested on our external cohort and evaluated using metrics such as recall, specificity, and F1 score.

Results

The training dataset included 141 CXR images. Of these, 30% (42 out of 141) displayed "focal" phenotypes, collected across three French university hospitals. After assessing multiple models, three were selected: the CheXFound model (both fine-tuned and original versions) and the RAD-DINO ViT-Base model (fine-tuned).

For the external dataset, comprising 165 CXR images, 40% (66 out of 165) of patients exhibited "focal" phenotypes. The final model selected was CheXFound (fine-tuned).

Discussion

For radiological phenotypes to be clinically actionable, their identification must be accurate and standardised. Image-based machine learning classifiers can help guide personalised ventilation strategies and reduce reliance on CT scanning. Our model has the potential to accurately identify these phenotypes, particularly in challenging cases that rely solely on CXR. The subsequent step will involve prospective validation and development of a user-interface to enable both clinicians and researchers to access phenotype information efficiently.

Advances in Multi-State Modeling

Wed, 30.09.26, 11:00–12:30, Room: N5

Chair: Fotios Siannis

Modelling time-restricted transitions in multi-state models: application to primary care referral pathways

11:00-11:15

Presenter: Georgy Gomon

Authors: Georgy Gomon; Ilaria Prosepe; Rachel Knevel; Saskia Le Cessie

Affiliations: Leiden University Medical Center, Leiden

Introduction

In primary care, multi-state model can be used to describe the diagnostic process of patients with musculoskeletal (MSK) complaints, capturing transitions between consultations, management decisions, and referrals to rheumatology. In this setting, transitions to referral typically occur either directly at the consultation or within a short period thereafter (e.g., when laboratory results become available). We refer to such transitions as time-restricted transitions, as these are transitions that can only occur within a specific time window after entering a state. Standard multi-state approaches that rely on the Markov assumption - such as using a clock-forward time scale - do not explicitly account for such time-restricted restrictions and may therefore yield biased estimates.

Methods

We compare four modelling strategies for time-restricted transitions in a multi-state setting: (i) a standard clock-forward (Markov) time-scale (ii) a clock-reset time-scale (iii) a clock-forward time-scale including time since state entry as covariate and (iv) a modified clock-forward method that constrains hazards to zero beyond a latent time window. The main estimand of interest are the transition probabilities between states. With a simulation study, we evaluate model performance under different data-generating mechanisms and covariate effects. We then showcase the impact of different modeling strategies in a large Dutch electronic health record registry of primary care MSK consultations.

Results

First simulation results show that, in the presence of covariates, the standard clock-forward approach yields biased transition probability estimates when transitions are time-restricted, regardless of the underlying data-generating mechanism. In contrast, the clock-reset approach seems to give unbiased results.

In the real-world dataset (2,358,750 primary care MSK episodes, 1,842,533 transitions, and 23,884 referrals), referrals predominantly occurred shortly after consultations, consistent with time-restricted behavior. The choice of modelling strategy substantially affected estimated referral probabilities.

Discussion

In the presence of time-restricted transitions, the commonly used clock-forward time-scale can lead to biased estimates of transition probabilities. Through simulation, we present insights into alternative modelling strategies and demonstrate that they can estimate transition probabilities without bias. In our real-world dataset, which exhibits time-restricted transitions, we observe substantial differences between approaches, underscoring the importance of appropriate time-scale selection.

Under-reporting of intermediate events in multi-state models

11:15-11:30

Presenter: Ilaria Prosepe

Authors: Ilaria Prosepe; Hein Putter; Mar Rodriguez-Girondo; Liesbeth de Wreede

Affiliations: Leiden University Medical Center, Leiden

Introduction

During a pandemic, it is valuable to study how individuals' trajectories through different states over time (illness, hospitalization, vaccination) influence mortality. Multi-state models provide a valuable tool for modeling trajectories. However, current estimation methods for multi-state models cannot account for under-reporting, which is common for illness and vaccination.

Methods

We consider the illness-death model, where individuals may transition directly to death or via illness. When the intermediate state (illness) is under-reported, naïve estimation approaches (e.g., Nelson-Aalen or partial likelihood + Breslow) yield biased estimates for most transition hazards. Only the illness-to-death hazard remains identifiable under the assumption that the illness-to-death hazard is the same for observed and unobserved illness, conditional on covariates.

We propose a novel plug-in + recalibration estimation method that incorporates an external source (either the hazard of death without illness or the hazard of illness) and demonstrate analytically that information on a single transition suffices to identify and reconstruct all transition hazards and transition intensities. This approach leverages the fact that, while illness is under-reported, death is fully observed or independently censored. We develop both non-parametric and semi-parametric (proportional hazards) versions of the estimation method and derive conditions under which all hazards are identifiable using analytical properties of multi-state models.

Results

We evaluate performance via simulations across six scenarios: no covariates, covariates, high under-reporting, small external data, and mild misspecification of external baseline hazards or coefficients. Naïve methods show substantial bias in all settings, especially with high under-reporting. In contrast, our approach accurately recovers cumulative hazards, coefficients, and transition probabilities when external data are well specified, and shows only minor bias under mild misspecification.

Discussion

In presence of under-reporting, naïve estimation methods for multi-state models lead to biased estimates of transition probabilities and transition hazards. With our proposed method, which integrates external information on one single transition, it is possible to substantially reduce bias when estimating transition probabilities and transition hazards. More broadly. the framework is applicable whenever risk-sets for some states are not reliable, including settings with under- or over-reporting.

Multistate Joint Modelling of Marker and Failure Data under Dual Censoring Schemes

11:30-11:45

Presenter: Laura Bumbulis

Authors: Laura Bumbulis; Richard Cook

Affiliations: University of Waterloo, Waterloo

Introduction

Interest often lies in the relation between a time-varying covariate (or 'marker') and a failure time. For example, in studies of prostate cancer, the relation between bone-specific alkaline phosphatase and the occurrence of skeletal complications (e.g. fractures) is often a focus of analyses. When longitudinal data are expensive to obtain, marker process follow-up often ends before the failure process is censored, resulting in missing marker values between the two censoring times.

Methods

We formulate a class of joint multistate models for the marker, failure, and random censoring processes that exploit the available marker data to learn from failure data observed after the marker censoring time, and we illustrate the utility of this framework in a discrete-time setting. We use an expectation-maximization algorithm for estimation with standard errors computed via Louis's method (1982), focusing on estimation of the failure model parameters. The asymptotic efficiency of the estimator from the proposed joint model using the data up to failure censoring is examined relative to the complete-data estimator that censors the failure process early upon marker censoring. We also investigate the sensitivity of the proposed estimator to various incorrect model assumptions, including time-homogeneity and misspecified fixed and random effects structures.

Results

The efficiency loss from early censoring can be substantial (over 90%) under certain joint processes, motivating consideration of joint modelling rather than standard failure time analysis. More efficiency is lost when the amount of missing data is large and/or the failure rate is high. Under the misspecified models, the percent relative asymptotic biases are modest (typically less than 5%) even for large amounts of missing data these biases are worse when the failure rate is high and/or the extent of the misspecification is great.

Discussion

The proposed class of joint multistate models can be used to substantially improve the efficiency of failure model parameter estimation when marker-failure data are collected under a dual censoring scheme. These models appear quite robust to various kinds of misspecification. While we focus on a relatively simple setting, the proposed framework accommodates additional features such as disease-dependent censoring, higher-order Markov transition intensities, intermittent observation, multiple categorical markers, and competing risks.

Natural history of disease modelling in multiple long-term conditions - multi-state modelling in chronic kidney disease, type-2 diabetes, and obesity

11:45-12:00

Presenter: Monil Khera

Authors: Monil Khera; James Rafferty; Timothy Osborne; Jane Lyons; Rhiannon K Owen

Affiliations: Swansea University, Swansea

Introduction

Understanding how coexisting conditions progress over time and their associated impact on patient outcomes is important to inform healthcare decision-making. This study investigated how obesity, type-2 diabetes (T2D), and chronic kidney disease (CKD), and their combinations, developed over time and how the temporal sequence of disease accumulation impacted life expectancy.

Methods

We used anonymised individual-level demographic, administrative and electronic health record data for the population of Wales from the Wales Multimorbidity e-cohort in the SAIL Databank. Our cohort included individuals aged 25 years and over who were living in Wales on Jan 1, 2000 with a 5-year clearance of obesity, T2D and CKD. Individuals were followed up from Jan 1, 2005 to Dec 31, 2019, death, or the first break in Welsh residency. Multi-state Cox proportional hazard models adjusted for age, sex and deprivation were used to analyse the natural history of disease and the associated impact on all-cause mortality.

Results

Our study included 1,368,910 individuals. We found that the order in which the three conditions were developed as well as the number of conditions developed had a significant impact on mortality risk. For example, CKD followed by obesity had the worst prognosis resulting in an estimated loss in life expectancy of 3.42 years (SD 0.03) however, if this was followed by a diagnosis of T2D this resulted in a slight gain in life expectancy (0.05 years (SD $\pm$0.00)) for a 50-year-old man in the most deprived community. We found that a significantly lower proportion of patients diagnosed with obesity and CKD in any order were prescribed multi-indication treatments such as metformin and Sodium Glucose Co-Transporter (SGLT2) inhibitors compared to patients that had an additional diagnosis of T2D ($p<0.001$ for all combinations).

Discussion

The development of T2D generally resulted in improved life expectancy, which may be associated with the use of multi-indication treatments. The introduction of multi-indication treatments such as metformin, SGLT2-inhibitors or glucagon-like peptide-1 (GLP-1) receptor agonists may introduce additional health inequalities related to diagnoses where these technologies are approved for use. Future work will explore the impact of multi-indication technologies in patient-level simulations.

Spectral Clustering of Parametric Multistate Models for Uncovering Patterns in Event History Data

12:00-12:15

Presenter/Author: Zoe Kristin Lange

Affiliations: Ruhr-University Bochum, Bochum

Introduction

Multistate models are widely used to analyze event history data documenting how individuals transition between clinically meaningful states over time. In practice, these data are rarely generated from a single homogeneous population, but patients are naturally grouped by healthcare provider, treatment strategy or clinically defined strata. Each group can exhibit its own transition dynamics, yet within-group information is often limited due to censoring and rare events.

This motivates identifying shared transition patterns across groups to enable cluster-level analysis. Clustering supports systematic comparisons between groups, improves statistical stability by pooling information across similar groups, and highlights outlying transition behavior for further investigation.

Methods

For each group, we fit a continuous-time multistate model with parametric transition intensities. We then construct a weighted similarity graph whose nodes correspond to the groups and whose edge

weights are given by a kernel function applied to a distance between fitted intensity functions. The clustering task is formulated as a graph partitioning problem where within-cluster similarity is high while between-cluster similarity is small, and can effectively be solved via spectral clustering with a normalized graph Laplacian matrix. Our proposed algorithm detects clusters in non-convex regions and does not require training data.

Results

Using the counting process framework and martingale concentration inequalities, we derive finite-sample bounds for the misclustered fraction of groups. We allow for the number of groups as well as the within-group sample size to grow and give conditions under which the algorithm is consistent, meaning that the misclustered fraction converges to 0.

Furthermore, the bounds clarify how clustering performance depends on the modeling parameters and show interpretable failure modes of clustering.

We construct an extensive simulation study across censoring regimes, sample sizes and cluster specifications to corroborate the theoretical results. We provide guidance for the choice of scaling parameters and the estimation of the true number of clusters.

Discussion

The proposed method provides a principled, theory-backed approach to cluster event-history processes in settings with many groups and limited within-group sample sizes. The similarity construction is adaptable to a broad range of scientific questions and the finite-sample theory offers practical guidance for study design and implementation.

Integrating Multi-State Modeling and graph-based clustering for longitudinal cancer subtyping

12:15-12:30

Presenter: Lara Cavinato

Authors: Lara Cavinato [1]; Marco Rocchi [1]; Francesca Ieva [1, 2]

Affiliations: [1] MOX - Modelling and Scientific Computing Lab, Department of Mathematics, Politecnico di Milano, Milano; [2] HDS - Health Data Science Centre, Human Technopole, Milano

Introduction

Cancer subtyping is pivotal for prognosis assessment and treatment selection, yet most existing approaches rely on static, biopsy-derived or cross-sectional biomarkers that fail to capture temporal disease evolution. Such strategies overlook heterogeneity emerging from treatment pathways and transitions across clinical states. We propose a trajectory-informed clustering model that integrates cross-sectional covariates with longitudinal disease dynamics through a multi-state modeling perspective. Our objective is to derive clinically meaningful cancer subtypes aligned with both phenotypic features and progression trajectories.

Methods

We construct a patient similarity graph combining (i) distances between baseline and imaging-derived covariates and (ii) distances between patients' cumulative transition risks across predefined clinical states (pre-surgery, chemotherapy, surgery, post-surgery, relapse, death), estimated via a multi-state model. Graph estimation, multi-state parameter estimation, and graph segmentation - resembling spectral clustering - are performed jointly through the optimization of a single penalized loss function. Regularization enables feature selection and scalability. We propose an alternating optimization algorithm to solve the resulting non-convex problem and establish theoretical convergence guarantees. Performance is evaluated through comprehensive simulations assessing predictive accuracy, latent stratum recovery, and scalability relative to state-of-the-art subtyping methods. An ablation study quantifies the contribution of joint learning and graph regularization. The framework is applied to a real-world dataset of patients with colorectal liver metastases, integrating CT-derived radiomic features (pre- and post-chemotherapy), surgical data, relapse, and survival outcomes.

Results

Simulation studies demonstrate that joint learning of graph structure and multi-state dynamics improves prediction of transition probabilities compared with sequential approaches. The graph-based term enhances recovery of latent subtypes, while regularization identifies relevant covariates and ensures computational scalability. Ablation analyses confirm that removing trajectory or graph components degrades clustering accuracy and prognostic discrimination. In the colorectal liver metastases cohort, the method identifies subgroups characterized by distinct radiomic profiles and significantly different relapse and survival trajectories, providing clinically interpretable subtype structures.

Discussion

Trajectory-informed clustering bridges static biomarkers and dynamic disease evolution by embedding multi-state risks into a graph-based framework. The proposed joint optimization approach improves subtype discovery, interpretability, and prognostic relevance. These findings highlight the potential of temporally aware, graph-regularized methods to advance personalized oncology.

14:00–15:30

Platform and Cluster Trial Designs

Wed, 30.09.26, 14:00–15:30, Room: K1

Chair: Diane Uschner

Sample Size Calculation for Hierarchical Composite Endpoints in Cluster Randomized Trials

14:00-14:15

Presenter: Hrishikesh Chakraborty

Authors: Hrishikesh Chakraborty; Sophia Waymyers

Affiliations: Duke University, Durham

Introduction

Hierarchical composite endpoints (HCEs) are designed to incorporate multiple clinically relevant outcomes into a single measure, prioritizing these outcomes based on their importance to both clinicians and patients. Cluster randomized trials (CRTs), which involve randomizing interventions at the group level rather than the individual level, are commonly employed when individual randomization is not feasible. The design and analysis of CRTs require careful consideration of intra-cluster correlation (ICC). Accurate estimation of sample size for CRTs utilizing HCEs is crucial to ensure sufficient statistical power while appropriately accounting for both clustering effects and the hierarchical nature of outcomes.

Methods

We propose a new approach for sample size estimation in CRTs employing HCEs. Composite endpoints were first transformed into score based ordinal outcomes, after which we developed sample size estimation formulas using a proportional odds model for clustered ordinal endpoints, with adjustments for ICC.

Results

We investigated how varying conditions - such as effect size, cluster size and variability, and ICC influence sample size requirements through simulation analysis. Additionally, we applied the developed methodology to two real-world trial datasets to demonstrate its utility: dialysis patient data from the HiLo trial and HIV study data from South Carolina.

Discussion

As interventional CRTs increasingly implement HCEs, the development of robust sample size estimation methodologies is essential for upholding rigorous research standards. Traditional approaches that overlook clustering or hierarchical structures tend to underestimate required sample sizes, thereby compromising validity. Our proposed method adheres to foundational principles of sample size determination by accounting for all sources of variability. It offers computational efficiency and flexibility across various CRT designs, delivering practical guidance for effective trial planning.

Power calculations for batched stepped wedge designs when treatment effect varies by batch

14:15-14:30

Presenter: Ehsan Rezaei-Darzi

Authors: Ehsan Rezaei-Darzi; Jessica Kasza

Affiliations: Monash University, Melbourne

Introduction

Stepped wedge designs are variants of longitudinal cluster randomised trials, and have been used with increased frequency in recent years. Batched stepped wedge designs are variants that allow for cluster recruitment as the trial is being conducted. Formulas for the statistical power of batched stepped wedge designs to detect effects are available, but these have assumed a constant treatment effect common to all batches. Such an assumption may be unrealistic and here we allow for the treatment effect to vary across batches. We consider the impact of this treatment effect heterogeneity on power and sample size calculations.

Methods

We consider linear mixed models for continuous outcomes with a random treatment effect for each batch and a random cluster-period effect, allowing for non-zero correlations between the random treatment effect and the cluster-period random effects. We derive an expression for the variance of the treatment effect estimator for batched stepped wedge designs based on this model, to be used in power and sample size calculations. The performance of this expression when used to generate required sample sizes is assessed using a simulation study, with a focus on situations with small numbers of batches. We illustrate the use of this variance expression with an example motivated by a stepped wedge trial in Australian rehabilitation units (the Disinvestment trial).

Results

As expected, when allowing for heterogeneity of the treatment effect across batches, larger sample sizes are required to detect effects. Using parameters inspired by the Disinvestment trial, when detecting an effect size of 0.2 the study power is reduced by 16% when treatment effect heterogeneity is present compared to the situation where there is no treatment effect heterogeneity. In

the simulation study, empirical power matches theoretical power when there are a large number of batches, but there is a discrepancy for small numbers of batches.

Discussion

The inclusion of a random treatment effect across batches increases the variance of the treatment effect estimator and reduces the power of the design. However, despite this increase in variance, such a model may be more realistic e.g. when some batches are in different geographical locations.

Group sequential batched stepped wedge design: an optimal design for stepped wedge trials with staggered cluster enrolment

14:30-14:45

Presenter: Richard Tawiah

Authors: Richard Tawiah; Andrew Forbes; Jessica Kasza

Affiliations: Monash University, Melbourne

Introduction

In a batched stepped wedge (SW) cluster randomised trial, clusters are enrolled in batches in a staggered fashion, with each batch composed of a SW design. Designs that allow for early termination of batched SW trials could thus be developed, potentially shortening study duration, and reducing the number of clusters recruited in a study. Group sequential designs have been used extensively in individually-randomised trials. They provide a sequence of interim analyses of a trial's data from which a decision to terminate the trial early for efficacy, harm or futility could be drawn.

Methods

In this talk, we will introduce the group sequential batched SW design, a unified framework that allows interim analysis with stopping rules in a batched SW trial, assuming continuous outcomes with a cross-sectional sampling scheme, and an outcome model with batch-specific period effects. We will consider designs in which interim analysis is conducted after each batch based on pre-planned interim monitoring schemes with equally spaced intervals induced by batched SW trials with identical batches.

Results

Theoretical properties of the group sequential batched SW design will be presented, showing that it has an independent increments covariance structure which thus allows the use of results for group sequential designs for individually-randomised trials. We will consider the error spending approach to the calculation of group sequential boundaries. Using simulation studies, we will show that our proposed design preserves type I error and power, and has the potential to reduce the expected total number of clusters and participants required. We will demonstrate the potential gains in sample size

through application of the proposed design to the EAGLE trial, a batched SW trial which included 355 hospital clusters in 16 batches.

Discussion

Batched SW trials may require a large number of clusters and take a long time to conduct. The potential gains in sample size and reductions in required time provided by the group sequential batched SW design mean that these trials can be completed sooner and for a lower cost.

Assessing the importance of model misspecification in the analysis of stepped-wedge trials with time-varying effects

14:45-15:00

Presenter: Marc Parsons

Authors: Marc Parsons [1]; Lawrence Mbuagbaw [2]; Shirin Golchi [1]; Guillaume Fontaine [1, 3]

Affiliations: [1] McGill University, Montréal; [2] McMaster University, Hamilton; [3] University of New South Wales, Sydney

Introduction

Stepped wedge trials (SWTs) are a subset of cluster randomized trials in which an intervention or treatment is successively randomized cluster-by-cluster over discrete time periods. These designs have logistical and ethical advantages over other designs in certain clinical contexts. They require careful statistical consideration, however, to account for time-varying intervention effects while controlling for correlation at multiple levels. Past work has highlighted shortcomings in published analyses of these studies at the trial level.

Evidence from SWTs may be synthesised in meta-analysis (MA). As in the analysis of single trials, MA models must also consider the issue of time-varying treatment effects and multi-level clustering. In practice, however, this is hampered by a lack of established MA methods for the analysis of SWTs.

Methods

We will investigate the effect of trial- and meta-level model misspecification on the MA of SWTs in the presence of time-varying intervention effects. Data will be first generated at the participant level following the closed-cohort SWT model of Hooper et al. (2016). In simulations, we will vary the presence and shape of a time-varying intervention effect. These data will be analyzed using either correct or misspecified time models as well as misspecified correlation models at the trial-level. The resulting estimates will then be further synthesized using either correct or misspecified models at the meta-level. The primary simulation parameter of interest is the proportion of correctly modelled trials included in the meta-analysis. Model performance will be measured using bias in intervention effects over time as well as coverage.

Results

Primary simulations (n=100 replications by scenario) indicate that both study- and meta-level model misspecification contribute to poor model performance in terms of bias and coverage. The magnitude of the performance effect appears to be larger for study-level models which are misspecified by time than the other misspecification scenarios.

Discussion

We expect to find in our full simulation study that model performance in the meta-analysis of SWTs may be more influenced by study-level factors than those at the meta-analysis level. These results will help inform researchers on the proper analysis and reporting of these types of trial design as well as their proper meta-analysis.

Multiple treatment arms, multiple biases? Allocation and chronological biases in rare disease platform trials

15:00-15:15

Presenter: Daniel Bodden

Authors: Daniel Bodden [1, 2]; Nico Bruder [3]; Franz Koenig [3]; Ralf-Dieter Hilgers [2]

Affiliations: [1] RWTH Aachen University, Aachen; [2] Sigmund Freud Private University Vienna, Vienna; [3] Medical University of Vienna, Vienna

Introduction

Innovative trial designs are increasingly needed for studies in small populations such as rare diseases, including paediatrics. Platform trials can improve efficiency by sharing infrastructure and control data, but staggered arm entry and operational flexibility may increase vulnerability to allocation bias and chronological bias, especially when non-concurrent controls are used.

Methods

Using Monte Carlo simulations of a three-arm platform trial (two experimental arms, one shared control), we compared complete randomization, permuted block randomization, and a generalized big-stick design under 1-step versus 2-step randomization. Allocation bias was modelled via a platform-trial extension of the Blackwell-Hodges convergence strategy, representing an unblinded recruiter who exploits predictable assignments chronological bias was introduced through time trends. Analyses used concurrent controls only or pooled concurrent and non-concurrent controls, applying a one-sided two-sample t-test or a period-adjusted linear model.

Results

Permuted block randomization with small blocks produced substantial type I error inflation under allocation bias, most pronounced under 2-step randomization complete randomization and the big-stick design were more robust. With concurrent controls only, time trends did not inflate type I error but reduced power. Pooling non-concurrent with concurrent controls under a linear time trend increased type I error a period-adjusted linear model controlled type I error.

Discussion

We recommend avoiding permuted block randomization with small blocks in unblinded trials and adjusting for time (e.g., period effects) when borrowing non-concurrent controls. Because 2-step randomization increased susceptibility to allocation bias, it should be planned cautiously in unblinded settings when feasible, 1-step randomization is a more robust default.

A Hybrid Randomisation Framework for Platform Trials with Delayed Reveal of Eligibility

15:15-15:30

Presenter: Rajenki Das

Authors: Rajenki Das; Sofía S Villar

Affiliations: MRC Biostatistics Unit, University of Cambridge, Cambridge

Introduction

Platform trials in intensive care and emergency medicine face a tension between statistical rigour and operational feasibility. While traditional "revealed" randomisation ensures balance, it often requires multiple, sequential randomisation points that are burdensome for clinical staff in time-critical settings. Conversely, a full "delayed reveal", where all assignments are made at patient enrolment but hidden until eligibility is confirmed, minimises staff burden but often leads to high allocation imbalance in intervention domains with latent eligibility.

Methods

We propose a middle-ground approach that balances operational simplicity with robust imbalance control. Our method utilises a delayed-reveal framework but incorporates a dynamic rebalancing mechanism. This system allows for a single enrolment point for clinical staff (minimising burden) while utilising periodic, data-driven "sequence rewrites" to adjust for observed eligibility rates. We evaluated this hybrid approach using a motivating example of a three-domain trial where the third domain is reserved for a small, unknown fraction of the population. We performed simulations to compare its performance against both static delayed-reveal and multi-point randomisation schemes.

Results

Simulations show that while a standard delayed-reveal protocol leads to large allocation imbalance, our proposed hybrid system effectively maintains balance within intervention domains and clinical strata. We found that the success of this middle ground approach depends on the frequency and logic of the sequence rewrites a naïve implementation can fail, but incorporating a dynamic rebalancing parameter ensures stable performance throughout the trial.

Discussion

Our findings illustrate that trialists do not have to choose between a burdensome multi-point system and a statistically risky delayed-reveal system. The proposed hybrid randomisation framework offers a practical solution for acute care platform trials, maintaining high statistical efficiency and balance while protecting the workflow of frontline clinicians.

Omics in Modeling

Wed, 30.09.26, 14:00–15:30, Room: K2

Chair: Niklas Brunn

Outcome-guided clustering within a generalised Bayes framework

14:00-14:15

Presenter: Sida Chen

Authors: Sida Chen; Sylvia Richardson; Brian Tom

Affiliations: University of Cambridge, Cambridge

Introduction

In molecular precision medicine, there is a growing interest in identifying disease subtypes from high-dimensional molecular-level covariate data that are associated with clinically relevant outcomes. A motivating example arises in proteomic studies of knee osteoarthritis, where the aim is to identify patient subgroups whose proteomic signatures may support better disease stratification and prediction, and eventually more targeted treatment or prognostic assessment. Bayesian profile regression is an established method for such outcome-guided clustering. However, it may suffer from the outcome being weakly relevant when informing the clustering structure and from poor out-of-sample outcome prediction in realistic settings potentially limiting its practical utility.

Methods

We propose a generalised Bayes (GB) framework for outcome-guided clustering by reformulating the problem as a multi-task decision problem with separate loss components (tasks) for clustering and prediction. The framework offers great flexibility and admits standard Bayesian profile regression as a special case. To improve the interpretability of the weighting of different loss components, we propose a principled standardisation that places each task risk on a common scale. Selection of the learning rate and task weights can be guided by an estimate of the expected out-of-sample risk under a user-chosen evaluation loss, which can be conveniently obtained from Markov chain Monte Carlo output.

Results

Preliminary simulation studies indicate that, with a suitably chosen evaluation loss and tuning parameters, the proposed GB approach achieves satisfactory cluster recovery and out-of-sample outcome prediction in challenging examples where standard profile regression fails. Results from an application to proteomic knee osteoarthritis data are also anticipated.

Discussion

Some pitfalls of standard Bayesian profile regression arise from inherent limitations of the likelihood-based specification. GB provides a decision-theoretic framework that offers the flexibility to encode explicit priorities across tasks and to deliver more goal-aligned performance, with potential for broader modelling and application settings.

Learning Latent Representations of Sepsis from Longitudinal Plasma Proteomics

14:15-14:30

Presenter: Kerstin Kappler

Authors: Kerstin Kappler [1, 2]; Thomas Schaaf [3, 4]; Malte Bayer [3, 4]; Britta Westhus [3]; Nadine Hennigs [1]; Katharina Rump [3]; Björn Koos [3]; Robin Grugel [4, 5]; Karin Schork [4, 5, 6]; Martin Eisenacher [4, 5, 6, 7]; Tim Rahmel [3]; Matthias Unterberg [3]; Lars Bergmann [3]; Katrin Marcus-Alic [4, 5]; Michael Adamzik [3]; Barbara Sitek [3, 4]; Thilo Bracht [3, 4]; Hartmuth Nowak [1, 3]

Affiliations: [1] Ruhr University Bochum, Knappschaft Kliniken University Hospital Bochum, Department of Anesthesiology, Intensive Care Medicine and Pain Therapy, Center for Artificial Intelligence, Medical Informatics and Data Science, 44892 Bochum, Germany, Bochum; [2] Institute for Technologies and Management of Digital Transformation (TMDT) University of Wuppertal, Wuppertal; [3] Ruhr University Bochum, Knappschaft Kliniken University Hospital Bochum, Department of Anesthesiology, Intensive Care Medicine and Pain Therapy, 44892 Bochum, Germany, Bochum; [4] Ruhr University Bochum, Medical Faculty, Medizinisches Proteom-Center, 44801 Bochum, Germany, Bochum; [5] Ruhr University Bochum, Medical Faculty, Medical Proteome Analysis, Center for Proteindiagnostics (PRODI), 44801 Bochum, Germany, Bochum; [6] Ruhr University Bochum, Medical Faculty, CUBiMed.RUB, Core Unit Bioinformatics, 44801 Bochum, Germany, Bochum; [7] Ruhr University Bochum, Medical Faculty, BioInfra.Prot, Deutsches Netzwerk für Bioinformatik-Infrastruktur - de.NBI, 44801 Bochum, Germany, Bochum

Introduction

Annually, ~280,000 cases of sepsis occur in Germany [1], representing a life-threatening systemic multi-organ failure. Treatment is largely limited to anti-infective and supportive measures, while targeted therapies remain unavailable. Molecular-mechanistic approaches have failed in clinical trials, potentially due to the syndromes complexity and pronounced heterogeneity of patient cohorts [2],[3]. With high-throughput technologies, detailed molecular characterization of patients has become feasible. Prior studies have identified subtypes based on plasma proteomics [4]. However, temporal dynamics of these molecular changes remain insufficiently explored. This study provides new insights into sepsis dynamics through analysis of longitudinal proteomic data.

Methods

We analyzed >1,500 plasma samples from 473 patients across two sepsis cohorts: SepsisDataNet.NRW (n = 231 day 1, 4, 8) and KI.SEP (n = 242 day 1-6, 8), using liquid chromatography-coupled mass spectrometry. For dimensionality reduction, a masked beta-variational autoencoder (VAE) was trained to project high-dimensional proteomic data into an eight-dimensional latent space. This latent space serves as a compact, continuous representation of patient-specific molecular states, enabling integration of heterogeneous signals and facilitating downstream analyses. For biological interpretation, SHAP values were computed. The latent representations were further analyzed using continuous-time hidden Markov models (CT-HMM) to characterize temporal dynamics, identify discrete disease states, and estimate transition probabilities.

Results

The latent model demonstrated stable performance on unseen patients. Specific latent dimensions appeared to capture consistent protein patterns, reflecting reproducible biological processes. Several distinct disease states characterized by characteristic proteomic signatures are identified. Temporal transitions between states during disease progression could be modeled, with the KI.SEP cohort revealing additional dynamics. The results are preliminary and are currently undergoing further validation.

Discussion

We suggest that VAE-derived latent representations provide a flexible framework to study sepsis at systemic level, capturing both discrete states and continuous disease trajectories. While CT-HMM can model temporal dynamics, the latent space supports a broader range of analyses of disease progression. Increased temporal resolution enables novel insights into disease dynamics and state transitions. This may contribute to improved characterization of sepsis trajectories and support the identification of biologically meaningful patterns and potential biomarkers.

References:

[1]doi: https://doi.org/10.1007/s00063-021-00777-5

[2]doi: https://doi.org/10.1038/nrdp.2016.45

[3]doi: https://doi.org/10.4161/viru.26516

[4]doi:0.1186/s13054-025-05639-6

Asymmetric Shapley values to quantify the importance of the genetic component in clinical prediction models

14:30-14:45

Presenter: Jeroen Goedhart

Authors: Jeroen Goedhart [1]; Mark van de Wiel [2]; Kjersti Aas [3]; Martin Jullum [3]

Affiliations: [1] Amsterdam University Medical Centers, Amsterdam; [2] Amsterdam University medical centers, Amsterdam; [3] Norwegian Computing Center, Oslo

Introduction

Assessing the added value of high-dimensional genomic variables beyond established clinical risk factors remains challenging in clinical prediction modelling. A common strategy is to compare the predictive performance of models fitted with and without the genomic component [1]. However, this approach can be misleading because it does not account for collinearity nor known directionality of dependencies between variables.

Methods

To address these issues, we propose to use asymmetric Shapley values [2,3] to quantify variable importance when clinical and omics variables are combined in prediction models. We focus on a clinically relevant setting in which disease progression, for instance measured by tumour stage, mediates part of the genomic effects, while additional clinical variables may act as confounders with unknown directionality. For this setting, we derive efficient algorithms to compute both local and global asymmetric Shapley values. Global Shapley values provide model interpretation by decomposing predictive performance metrics into contributions of the individual variables. For the more commonly used local Shapley values, we develop an inferential procedure that enables statistical testing of the relevance of high-dimensional omics variables.

Results

We illustrate the proposed framework using a case study on the prediction of progression-free survival in colorectal cancer patients, based on tumour stage, established clinical risk factors, and high-dimensional gene expression data. Compared with conventional approaches, our framework attributes a larger contribution to the genomic component. Furthermore, our framework reveals an interesting interaction effect between the genes and tumour stage.

Discussion

These findings illustrate that incorporating directional dependencies between predictors can substantially change how the predictive value of high-dimensional omics data is quantified and interpreted in clinical prediction models.

References

[1]: De Bin R., Sauerbrei W. and Boulesteix A.L. (2014), Investigating the prediction ability of survival models based on both clinical and omics data: two case studies, Statistics in Medicine, 33, 5310-5329

[2]: Kjersti Aas, Martin Jullum, Anders Løland (2021), Explaining individual predictions when features are dependent: More accurate approximations to Shapley values, Artificial Intelligence, 298, 103502,

[3] Frye, C., Rowat, C., and Feige, I. (2020), Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability. Advances in neural information processing systems, 104, 1229-1239.

Two-stage rank-based identification of surrogate signature from high-dimensional data applied to vaccinology transcriptomics

14:45-15:00

Presenter: Boris Hejblum

Authors: Arthur Hughes [1]; Layla Parast [2]; Rodolphe Thiebaut [3]; Boris Hejblum [4]

Affiliations: [1] INSERM, INRIA, BPH, U1219, SISTM, University of Bordeaux, Vaccine Research Institute, Bordeaux; [2] Department of Statistics and Data Science, University of Texas at Austin City: Austin (TX), Austin; [3] Université de Bordeaux, Inserm, Bordeaux Population Health Research Center, U1219, Bordeaux; [4] INSERMBordeaux Population Health, Talence

Introduction

Measuring clinically relevant endpoints in vaccine trials can be costly, time-consuming, or impractical, making surrogate markers valuable tools for indirectly capturing treatment effects. High-throughput technologies such as RNA-sequencing offer promising candidates for surrogate identification, but generate high-dimensional data that pose unique statistical challenges. Existing surrogate marker methods are ill-suited to small-sample, high-dimensional settings typical of modern vaccine experiments, and those operating within a single-trial framework rely on assumptions about external generalisability that limit their broader applicability.

Methods

We introduce RISE (Rank-based Identification of high-dimensional SurrogatE Markers), a novel approach designed for small sample, high-dimensional settings typical in modern vaccine experiments. RISE is a non-parametric parametric leveraging two one-sided tests to perform univariate screening of each variable for promising surrogate candidates, followed by an evaluation of their weighted combination with internal or ideally external validation. We further extend this approach for meta-analytic evaluation of surrogate markers across multiple trials, relying on random-effect model to derive a combined surrogate signature using weights reflecting both surrogacy strength and cross-study variability.

Results

Our simulation studies demonstrate RISE's desirable properties, confirming that RISE controls the type one error rate and achieves satisfactory empirical power under a wide range of conditions. Applied to a single inactivated influenza vaccine trial, RISE identified a transcriptomic signature serving as a reasonable surrogate for the neutralising antibody response. RISE-meta was subsequently applied to five inactivated influenza vaccine trials with paired antibody and transcriptomic measurements collected pre- and post-vaccination, identifying a day-1 post-vaccination surrogate signature that is predictive of the treatment effect on subsequent antibody responses across trials.

Discussion

Our results consistently highlight innate immune-related pathways, particularly innate antiviral signalling and interferon stimulation, providing a coherent immunological interpretation. By leveraging treatment effect heterogeneity across trials, RISE-meta additionally assesses surrogate transportability. This constitutes a new principled and scalable framework for high-dimensional surrogate marker identification with applications in immunology and vaccine research.

Diagnostic Modeling of 21-Hydroxylase Deficiency Based on Targeted GC-MS Urinary Steroid Metabolomics and Machine Learning

15:00-15:15

Presenter: Huijie Li

Authors: Huijie Li [1]; Thomas Remer [2]; Michaela Hartmann [1]; Stefan Wudy [1]; Jörn Pons-Kühnemann [3]

Affiliations: [1] Justus Liebig University Giessen, Giessen; [2] University of Bonn, Bonn; [3] Justus-Liebig Universität GießenInstitut für Medizinische Informatik, Gießen

Introduction

21-hydroxylase deficiency (21-OHD) is the most common cause of congenital adrenal hyperplasia (CAH). This study leverages high-dimensional GC-MS data and machine learning (ML) to develop a standardized diagnostic framework, with a specific focus on optimizing data preprocessing for enhanced visualization and classification.

Methods

This study retrospectively analyzed a cohort of 221 urinary samples from 21-OHD patients (aged >2 years mean age: 9.71 years) collected between 2001 and 2025, alongside 400 healthy control samples. Targeted GC-MS quantified 36 steroid metabolites. Different preprocessing techniques were applied to optimize the input feature space. The dataset was partitioned (70% training, 30% testing) using stratified sampling. A recursive partitioning-based decision tree was implemented to identify optimal diagnostic thresholds.

Results

Quantile normalization clearly improved heatmap readability by reducing the influence of extreme metabolic outliers compared to other standard normalization methods. The diagnostic model derived from the training set identified two biomarkers as the most critical predictors for 21-OHD diagnosis: 11-oxo-pregnanetriol ($\geq$84 μg/L) emerged as the primary predictor, followed by 17α-OH-pregnanolone ($\geq$79 μg/L) as secondary predictor. In the independent testing set, the model demonstrated robust diagnostic performance (Sensitivity: 91%, Specificity: 99%). A receiver operating characteristic curve was generated using the predicted probabilities from the testing set, the area under the curve reached an impressive 0.96 (95% CI: 0.92-0.99).

Discussion

While serum 17-OHP via LC-MS/MS is common for 21-OHD screening, its clinical utility is hindered by significant inter-laboratory variability. In contrast, integrating targeted GC-MS profiling with computational intelligence yields a highly stable diagnostic tool for 21-OHD. Our findings suggest that selecting appropriate statistical normalization is critical for the visual interpretation of complex steroidomes. This synergy between analytical chemistry and advanced statistics provides a scalable model for the precision diagnosis of adrenal disorders.computational intelligence could serve as a powerful tool for advancing precision medicine in clinical endocrinology.

AttenGene - a source-based, privacy-compliant AI assistant for clinical genetics using retrieval-augmented generation

15:15-15:30

Presenter: Oleg Borisov

Authors: Oleg Borisov; Matthias Wuttke; Anna Köttgen; Ulla T Schultheiss

Affiliations: University Medical Center Freiburg, Institute of Epidemiology and Prevention, Freiburg

Introduction

Large language models (LLMs) can assist in clinical genetics by synthesizing information across databases for variant interpretation. However, cloud-hosted LLMs raise privacy concerns when processing patient-related queries, and may generate unsupported statements (hallucination). We are developing AttenGene, an AI assistant that addresses both issues: Retrieval-Augmented Generation (RAG) grounds responses in cited source documents, while locally deployed compact models keep all processing on-premise.

Methods

AttenGene integrates genetics databases into a single query interface. Unstructured knowledge sources (e.g., GeneReviews) are embedded using BGE-large-en-v1.5 and retrieved via semantic search (PostgreSQL/pgvector, HNSW indexing). Structured databases (e.g., ClinVar) are queried via lexical search using trigram similarity. Retrieved documents are injected into the LLM context - Qwen3-14B (4-bit quantized) - to generate cited responses (>10 tokens/s). Guardrails enforce source attribution and suppress unsupported claims. The system is deployable on a single GPU workstation within institutional infrastructure.

Results

In natural language queries with gene and variant notation (e.g., "Is PKD1 c.6658C>T pathogenic?"), lexical search correctly retrieved the matching ClinVar record as top result in >=91.7% of cases (Recall@5: >=99.7%) across 326,397 variants (n=5,000 queries). For phenotype-based queries, semantic retrieval identified the correct GeneReviews disease chapter in 88.2% (n=187 queries, 51 conditions). Fabricated variant queries (n=1,000) were rejected at 96.5% with zero false rejections of true

variants. A pilot survey of genetics professionals (n=11) showed 91% daily database use with >30 minutes per case, and 91% rated having access to such a tool as "very helpful."

Discussion

Local, RAG-based AI systems provide viable, privacy-compliant clinical assistance grounded in curated genetic databases. AttenGene demonstrates that compact, locally deployed LLMs can achieve high retrieval accuracy without cloud dependency. This work was supported by the German Research Foundation (DFG) Collaborative Research Center (SFB) 1453 - Nephrogenetics (project ID 431984000).

Causal Estimands and External Controls

Wed, 30.09.26, 14:00–15:30, Room: N1

Chair: Vanessa Didelez

Adaptive estimators for the net benefit of a treatment

14:00-14:15

Presenter: Tomas Maiuri

Authors: Tomas Maiuri [1]; François Petit [1]; Antoine Chambaz [2]

Affiliations: [1] Université Paris Cité, Université Sorbonne Paris Nord, INSERM, INRAE, Centre for Research in Epidemiology and StatisticS (CRESS), Paris, France, Paris; [2] Université Paris Cité, Paris

Introduction

Assessing the effectiveness of a treatment is a central challenge in epidemiology and causal inference. While the average treatment effect (ATE) remains the most common estimand, medical decisions typically involve multiple criteria - to simultaneously assess benefit, risk, and tolerability for example - that a single outcome measure cannot capture. Measures based on hierarchical criteria, and in particular the net benefit [1], offer a well-motivated alternative by integrating these criteria through pairwise comparisons between treated and control subjects.

Methods

In this work, we rigorously formulate the net benefit within the potential outcomes framework and derive the efficient influence function (EIF) of the corresponding statistical parameter - a key object for constructing semiparametrically efficient estimators. To the best of our knowledge, this result completes a partial identification proposed in earlier work [2], and extends to the broader class of treatment effect measures based on pairs of observations. Building on this EIF, we propose an Augmented Inverse Probability Weighting Estimator (AIPWE) for the net benefit. This estimator enjoys two key guarantees: asymptotic efficiency under correct model specification, and double robustness when nuisance models are misspecified. We also introduce an alternative estimation strategy that bypasses the analytical difficulties arising from the pairwise structure of the estimand, at the cost of one additional estimation step.

Results

The theoretical properties of both approaches are established. Numerical experiments across a range of scenarios are underway to assess their finite-sample behavior, with particular attention to the impact of the nuisance parameter estimation strategy - comparing naive empirical methods, classical machine learning, and density estimation approaches. We also demonstrate their application to real data using the IST-3 dataset.

Discussion

A key limitation lies in specifying the hierarchical ordering of outcomes, which is a matter of debate and can affect the estimates.

References

[1] Marc Buyse. "Generalized pairwise comparisons of prioritized outcomes in the two-sample problem". In: Statistics in Medicine 29.30 (2010), pp. 3245-3257.

[2] Lu Mao. "On causal estimation using U-statistics". In: Biometrika 105.1 (2018), pp. 215-220.

Longitudinal time-in-range estimands for continuous outcomes under truncation by death

14:15-14:30

Presenter: Thomas Alexander Gerds

Authors: Thomas Alexander Gerds; Anders Munch

Affiliations: Section of Biostatistics, Department of Public Health, University of Copenhagen, Copenhagen

Introduction

Randomized trials and emulated target trials with continuous biomarker outcomes such as HbA1c or body mass index are challenged by truncation by death. Current ICH E9(R1) guidelines recommend defining a clinically meaningful estimand aligned with the clinical question and estimable under transparent assumptions.

Methods

We define a multi-state model framework with intermediate states based on biomarker values and suitably chosen thresholds, and with death explicitly included as an absorbing state. We define an estimand for the comparison on treatment regimens based on the percentage of expected lifetime spent in range within the next t-years. We also consider time-to-event estimands defined as probability of first entry into a target range of the continuous outcome, optionally requiring confirmation by a second measurement.

Results

We apply a longitudinal targeted minimum loss estimator accommodating right censoring and time-varying confounding to illustrate the methods using data from the LEADER trial and Danish nationwide registry data. In addition to specifying thresholds that define the range of the continuous outcome and an interpolation strategy for the biomarker trajectory between observation times, the estimation procedure requires mapping observations made in continuous time onto a discrete time scale.

Discussion

The clinical interpretation of the estimands remains dependent on differences in all-cause mortality between treatment regimens.

Revealing the Truth: Calculating True Values in Causal Inference Simulation Studies via Gaussian Quadrature

14:30-14:45

Presenter: Alex Ocampo

Authors: Alex Ocampo [1]; Enrico Giudice [2]; Zachary McCaw [3]; Tim Morris [4]

Affiliations: [1] Roche, Basel; [2] Novartis Pharma AG, Basel; [3] University of North Carolina at Chapel Hill, Chapel Hill; [4] Novartis, London

Introduction

Simulation studies are used to understand the properties of statistical methods. A key luxury in many simulation studies is knowledge of the true value (i.e. the estimand) being targeted. With this oracle knowledge in-hand, the researcher conducting the simulation study can assess across repeated realizations of the data how well a given method recovers the truth. In causal inference simulation studies, the truth is rarely a simple parameter of the statistical model chosen to generate the data. Instead, the estimand is often an average treatment effect, marginalized over the distribution of confounders and/or mediators.

Methods

The variables marginalized over are often generated from common distributions such as the normal, uniform, exponential, or gamma. For all these distributions, Gaussian quadratures provide efficient and accurate calculation for integrands with integral kernels that stem from known probability density functions. We demonstrate through three applications how to use Gaussian quadrature to accurately and efficiently compute the true causal estimand: covariate adjustment using g-computation, estimation of the controlled direct effect (hypothetical estimand), and mediation analysis with RMST. We also compare the pros and cons of Gauss–Hermite quadrature to Monte Carlo integration approaches, which we use as benchmarks.

Results

Gaussian Quadrature outperforms Monte Carlo based simulations of the truth in all cases. It is often instantaneous computationally and simulation based methods converge towards the truths computed by Gaussian quadrature. Some Monte Carlo integration computations have non-negligible bias that would propogate into bias and coverage probability operating characteristics of the simulation study, while others perform similarly to quadrature.

Discussion

Overall, we demonstrate that the Gaussian quadrature is an accurate tool with negligible computation time, yet is underused for calculating the true causal estimands in simulation studies. Utilizing Gaussian quadrature can ensure that the conclusions drawn from simulation studies are not subject to any Monte Carlo error.

Distribution shift as violations of key causal assumptions: implications for augmented clinical trial ATE estimation

14:45-15:00

Presenter: Maylis Tran

Authors: Maylis Tran [1]; Juliette Ortholand [2]; Sophie Tezenas du Montcel [1]

Affiliations: [1] Sorbonne University, Paris Brain Institute (ICM), INSERM, INRIA, CNRS, APHP, Paris; [2] Department of Medical Informatics, Amsterdam UMC, University of Amsterdam, Amsterdam

Introduction

Reducing sample size while preserving statistical power in randomized clinical trials (RCT) remains a central challenge, driven by cost, feasibility, and ethical concerns. Several methods leverage external observational data to improve efficiency, including Prognostic Covariate Adjustment (PROCOVA) or Prediction-Powered Inference for Clinical Trials (PPCT). These approaches incorporate prognostic scores derived from predictive models trained on observational data, into the Average Treatment Effect (ATE) estimation. Distribution shifts, defined as differences between observational and RCT data, may affect their efficiency. While the causal inference literature studies the assumptions required to use observational data to generalize RCT results without compromising validity, existing efficiency-enhancing methods are known to preserve validity regardless of the training data. However, much less is known about the conditions under which such data effectively improve efficiency.

We studied this problem through a causal inference perspective, linking forms of distribution shift to classic causal assumptions violations and identifying when using observational data improves efficiency.

Methods

We studied how different forms of distribution shift between the observational and RCT data relate to violations of classic causal assumptions. We simulated data to evaluate how these different forms of distribution shift, affect ATE estimators that integrate observational data compared to the basic difference-in-means (DM) one.

Results

We argue that distribution shifts harming efficiency correspond directly to violations of positivity (support inclusion) and unmeasured confounders. When no key assumptions were broken and only selection bias on observed covariates was present (scenario 1), the PPCT estimator reduced variance by 56% compared to DM. Under positivity violation (scenario 2), this gain dropped drastically to 0.53%. In the presence of an unmeasured confounder (scenario 3), the variance reduction dropped to 32%, reflecting model misspecification. Bias remained small across all assumption violations: -3.02% (DM) vs -3.66% (PPCT, scenario 1), 3.46% vs 3.50% (scenario 2), 1.26% vs 0.73% (scenario 3), confirming the previously stated validity preservation. We observed similar performances for the PROCOVA estimator. The impact of distribution shifts on real data from Amyotrophic Lateral Sclerosis disease also confirmed these results.

Discussion

Overall, efficiency gains from observational augmentation depend strongly on data alignment, reflecting positivity and the absence of unmeasured confounders.

Dynamic Borrowing and Prognostic Covariate Adjustment for RCTs With External Controls: A Neutral Comparative Simulation Study

15:00-15:15

Presenter: Luzia Berchtold

Authors: Luzia Berchtold [1]; Floor Van Oudenhoven [2]; Maarten van Smeden [1]; Joost van Rosmalen [1]

Affiliations: [1] University Medical Center Utrecht, Utrecht; [2] Danone Research & Innovation, Utrecht

Introduction

Randomized controlled trials that use information from external or historical controls aim to improve estimation efficiency but may suffer from bias or inflated type I error rates. Two prominent strategies are dynamic borrowing, based on outcome exchangeability, and prognostic covariate adjustment (PROCOVA), which leverages the transportability of covariate-outcome relationships. Despite their differences in assumptions, both approaches are proposed for similar trial settings, yet neutral and direct comparisons of their relative performance remain limited.

Methods

We conduct a neutral simulation study comparing common implementations of both strategies under a range of controlled data-generating mechanisms. Scenarios reflect trial incompatibility through systematic mean shifts, covariate distribution drift, and changes in covariate-outcome relationships, capturing possible lack of exchangeability and transportability. We compare a non-borrowing AN(C)OVA benchmark against PROCOVA, Bayesian and frequentist dynamic borrowing, and hybrid approaches. Performance is assessed using bias, mean squared error, type I error, power, and relative efficiency.

Results

Initial results confirm that dynamic borrowing is vulnerable to type I error rate inflation under data incompatibility, whereas PROCOVA maintains nominal error control, with misspecification of the prognostic model only affecting efficiency gains. For PROCOVA, efficiency gains are relative to the current trial sample size, whereas for dynamic borrowing, larger external sample sizes more directly translate into higher power, assuming data compatibility. Hybrid approaches approaches combining covariate adjustment with dynamic borrowing demonstrate the highest potential for efficiency gains.

Discussion

This study aims to provide an objective assessment of the strengths and limitations of these strategies and to inform methodological choice rather than advocate for a particular approach. The nature of the mismatch between external and current data appears critical for method selection, although such mismatches are often difficult to diagnose in practice. Hybrid approaches may offer a pragmatic compromise when some compatibility can be assumed.

A simulation study on the performance of causal inference approaches for external control arms with time-to-event endpoints in small-sample settings and under violation of proportional hazards

15:15-15:30

Presenter: Matthias Monnereau

Authors: Matthias Monnereau [1, 2]; Louise Baschet [2]; Stefan Michiels [1]; Thomas Filleron [1, 3]

Affiliations: [1] Université Paris-Saclay, CESP, INSERM U1018 Oncostat, labeled Ligue Contre le Cancer, Villejuif; [2] Horiana, Bordeaux; [3] Biostatistics & Health Data Science Unit. Oncopole Claudius Regaud, Toulouse

Introduction

External control arms (ECAs) are increasingly used in clinical development in oncology or rare diseases. The performance of causal inference methods in small samples with unmeasured confounding (UC) and non-proportional hazard (NPH) remains understudied for time-to-event endpoints. This study evaluates three methods: Standardized Mortality Ratio Weighting (SMRW), G-Computation (GC), and their combination (GC+SMRW).

Methods

We conducted Monte Carlo simulations mimicking a realistic ECA in non-small cell lung cancer. Survival times were generated using a piecewise exponential model with a delayed treatment effect (HR=1 until 3 months, HR=0.6 after). Scenarios varying sample size in the experimental arm (30, 60, 100), allocation ratio experimental: standard-of-care (1:1, 1:3), presence of UC with three confounding influences (none, weak, strong), and a null treatment effect scenario were investigated.

Treatment effects were measured using the difference of Restricted Mean Survival Time (Delta RMSTtau) with true Delta RMST=12.2 days at tau = 9 months and 52.3 at tau = 18 months for the average treatment on the treated population. Performance metrics included power/type 1 error, bias, accuracy and confidence interval coverage. SMRW uses a Kaplan-Meier estimator to compute RMST GC uses a Cox outcome model to reconstruct survival curves from which RMST is derived.

Results

All methods controlled type 1 error. SMRW outperformed other approaches in bias and coverage. GC and GC+SMRW exhibited a systematic overestimation of the treatment effect across scenarios with GC (range in days): Delta RMST9m=16.9-17.7 Delta RMST18m=59.3-61.3 and GC+SMRW: Delta RMST9m=15.9-18.1 Delta RMST18m=55.3-62.5. GC achieved the highest power. GC+SMRW presented intermediate performance. UC introduced a systematic bias in all three methods with no clear advantage for any method.

Discussion

Under NPH, GC and GC+SMRW gained power at the cost of bias from imposing the proportional hazards assumption in the outcome model, whereas SMRW provided unbiased but less powerful estimates. When RMST is the target estimand in ECAs with suspected NPH, flexible outcome models for GC should be explored.

Real-World Evidence, Registries, and Surveillance

Wed, 30.09.26, 14:00–15:30, Room: N2

Chair: Suzanne Cadarette

Evaluation of the Broad Consent Distribution at a University Hospital: Is it Representative or Biased?

14:00-14:15

Presenter: Tobias Brix

Authors: Justin Steven Hoffmann [1]; Philipp Lenz [2, 3]; Antje Duda [3]; Annalen Bleckmann [3, 4]; Nicole Niehoff [5]; Lena Clever [1]; Dominik Heider [6]; Tobias Brix [6]

Affiliations: [1] Department of Information Systems, University of Münster, Münster; [2] Department of Palliative Care, University Hospital Münster, Münster; [3] West German Cancer Center, University Hospital Münster, Münster; [4] Department of Medicine A (Hematology, Hemostasiology, Oncology and Pulmonology), University Hospital Münster, Münster; [5] Central Biobank of the Medical Faculty of Münster, University of Münster, Münster; [6] Institute of Medical Informatics, University of Münster, Münster

Introduction

Within the German Medical Informatics Initiative, the Broad Consent (BC) enables the secondary use of routine clinical data and biospecimens but raises concerns regarding representativeness and selection bias. This study evaluates BC coverage, consent rates, and representativeness of oncologic patients with a documented BC at a university hospital.

Methods

An anonymized dataset from the University Hospital Münster covering the period from September 2021 to September 2025 was analyzed. The study population comprised 38819 adult oncology patients, of whom 4835 had a documented BC. Representativeness was assessed by comparing the BC sample with the overall population using Chi-square goodness-of-fit tests and Cohen's w across demographic and clinical characteristics.

Results

The overall BC coverage rate was 12.46%, with higher coverage among stationary patients (17.74%) compared to ambulatory patients (9.56%). Among the documented BC cases, consent rates were high, with 96.27% of BC patients agreeing and 3.73% declining. BC module specific consent was consistently high, except for lower acceptance of third-country transfer options. All examined attribute distributions showed statistically significant differences. However, effect sizes varied: gender (w = 0.07) and age groups (w = 0.26) indicated small deviations, suggesting good representativeness. In

contrast, case status (w = 0.44) and ICD-10 diagnosis groups (w = 0.86) revealed moderate to large discrepancies.

Discussion

These results indicate limited representativeness of the BC sample, particularly for clinical characteristics, and suggest potential selection bias. Uneven BC implementation across clinical departments is a likely contributing factor. Future evaluations should reassess these effects after full implementation of BC to improve representativeness and support valid secondary data use.

The full paper of this abstract will be published in the GMDS series "Studies in Health Technology and Informatics" with doi: 10.3233/SHTI261012, URL:https://dx.doi.org/10.3233/SHTI261012

Leveraging linked real world health and social care data to evaluate the efficacy of home care services after hospital discharge in Lombardy, Italy

14:15-14:30

Presenter: Katherine Logan

Authors: Katherine Logan [1, 2]; Valeria Nazzari [1, 2]; Emanuele Di Angelantonio [3, 4, 5, 6, 7]; Francesca Ieva [8, 2]

Affiliations: [1] MOX - Modelling and Scientific Computing Lab, Department of Mathematics, Politecnico di Milano, Milano; [2] HDS - Health Data Science Centre, Human Technopole, Milano; [3] HDS - Health Data Science Centre, Human Technopole, Milan; [4] British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge; [5] Victor Phillip Dahdaleh Heart and Lung Research Institute, University of Cambridge, Cambridge; [6] British Heart Foundation Centre of Research Excellence, University of Cambridge, Cambridge; [7] National Institute for Health and Care Research Blood and Transplant Research Unit in Donor Health and Behaviour, University of Cambridge, Cambridge; [8] MOX lab, Department of Mathematics, Politecnico di Milano, Milan

Introduction

Globally, aging populations are exerting pressure on health and social services, highlighting the importance of effective alternative models of care, such as home care (HC), which seeks to reduce health system resource use and prevent poor patient outcomes [1]. In Italy, where the population aged 65 and over now represents ~25% of the population [2], these services are understudied [3]. This study leverages linked Italian health, and social care data to evaluate the effect of HC provision following hospitalization.

Methods

A multi-state Markov model was developed to simulate the patient outcomes (mortality) and health system resource use (emergency re-admission) of patients in Lombardy in the six months following a hospital discharge with the intention to initiate HC. Using propensity score matching, individuals who initiated HC within one week of discharge from hospital were matched to a cohort of individuals who did not.

Results

Following propensity score matching, an evenly matched cohort of 2,520 individuals was identified. Over the 180 days following discharge from hospital, we observed a higher rate of emergency hospital re-admission and lower mortality rate in the group that received HC. This difference extrapolates to 32 preventable deaths (3%) if every individual who needed HC after a hospitalization (as indicated on their discharge record) had received services in 2022.

Discussion

This study is the first to validate in an Italian setting the efficacy in preventing mortality of provision of HC following hospitalization. We interpret the increased risk of emergency re-admission as an indication of appropriate help seeking, possibly facilitated by increased contact with care professionals. Further analysis by type of hospitalization is useful for policymakers interested in effectively expanding these services.

References

[1] Kjerstad E, Tuntland HK. Reablement in community-dwelling older adults: a cost-effectiveness analysis alongside a randomized controlled trial. Health Economics Review. 2016 May 106(1).

[2] Population ages 65 and above (% of total population) - Italy. World Bank Open Data. 2024.

[3] Genet N, Boerma WG, Kringos DS, Bouman A, Francke AL, Fagerström C, et al. Home care in Europe: a systematic literature review. BMC Health Services Research. 2011 Aug 3011(1).

Application of Targeted Maximum Likelihood Estimation in a Target Trial Emulation: Investigating the Real-World Effect of Empagliflozin in Type 2 Diabetes Mellitus

14:30-14:45

Presenter: David Ryan

Authors: David Ryan [1]; Ruth Keogh [2]; Patrick Bidulka [2]; Anoop Shah [1]; Karla Diaz-Ordaz [1]; R T Lumbers [1]; Elizabeth Williamson [2]

Affiliations: [1] University College London, London; [2] London School of Hygiene and Tropical Medicine, London

Introduction

Target trial emulation (TTE) is increasingly used in pharmacoepidemiology to estimate drug effects using observational data. Confounding is typically addressed using propensity score (PS) or regression-based methods. Causal machine learning (ML) approaches offer greater modelling flexibility, including capturing non-linearities and complex interactions without pre-specification. However, their use in practice remains limited. This study applies targeted maximum likelihood estimation (TMLE), a doubly robust causal ML method, alongside conventional approaches, and explores challenges of applying TMLE using real-world data.

Methods

We re-analysed a published TTE study of the EMPA-REG RCT, comparing all-cause mortality between individuals with type 2 diabetes initiating empagliflozin versus DPP-4 inhibitors using UK primary care data (N = 12,970) [1]. We estimated marginal 3-year risk differences, risk ratios, and numbers needed to treat (NNT). Conventional approaches included inverse probability of treatment weighted Cox models and g-computation. TMLE was implemented with nuisance functions (PS, outcome and censoring models) estimated using either: (1) superlearner ensemble or (2) logistic regression. Covariate balance was assessed using standardised mean differences. Results were compared with the RCT.

Results

Results from conventional and TMLE analyses were directionally consistent with the RCT, indicating a protective effect of empagliflozin. TMLE produced narrower confidence intervals than conventional approaches. Under multiple imputation, TMLE with superlearner yielded a 3-year RR of 0.81 (95% CI: 0.75, 0.87) and an NNT of 24 (18, 37), while TMLE with logistic regression yielded an RR of 0.68 (95% CI: 0.56, 0.83) and an NNT of 14 (10, 26). Results from the missing indicator analysis were similar. Covariate balance was superior using logistic regression-based PS compared to super learner, with evidence of greater positivity violations under the latter.

Discussion

TMLE replicated the findings of an RCT in a TTE setting and relies on weaker assumptions regarding model specification for valid point estimation, which may be advantageous in real-world applications. However, practical challenges remain, including increased computational burden and complexities in handling missing data. Whilst TMLE is often pragmatically combined with multiple imputation, this raises unresolved methodological issues around congeniality and application of Rubin's rules to non-likelihood-based estimators.

Ryan DK et al. BMJ Open Diabetes Res Care. 2026 DOI

Non-Randomized Studies for treatment comparisons using German Registry Data: Insights from the NANA Project

14:45-15:00

Presenter: Enes Malik Cakir

Authors: Paula Starke; Enes Malik Cakir; Maxi Schulz; Susanna Salem; Tim Mathes; Tim Friede

Affiliations: Department of Medical Statistics, University Medical Center Göttingen, Göttingen

Introduction

The evaluation of therapies is a complex and time-consuming process, relying on randomized controlled trials (RCTs) as the gold standard. The increasing availability of digital health data has led to a growing interest in using real-world evidence for therapeutic decision making, particularly in rare diseases or time-critical questions where RCTs are impractical. However, the conditions under which such non-randomized studies using the target trial emulation framework produce valid and reliable results remain a subject of debate.

Methods

In this project, we conduct 10-12 non-randomized studies for comparing the benefits or risks of two alternative treatments with data from different German patient registries using state-of-the-art causal inference methodology and statistical modelling. We developed a staged workflow that is applied to all case studies: (1) pre-assessment of registry suitability (2) definition of a target trial in cooperation with clinical experts (3) identification of a minimally sufficient confounder adjustment set through a structured literature search and a clinician survey (4) specification of a detailed study protocol that accounts for data availability and (5) comparative analysis.We use targeted maximum likelihood estimators (TMLE) with flexibly estimated nuisance functions to protect against model misspecification, and apply diagnostics to estimated propensity scores. After conducting the studies, we compare the results with existing randomized controlled trials on the same clinical questions and analyse potential differences in the estimates and their causes.

Results

First results, methodological challenges and learnings from several completed case studies will be presented. We will also introduce a statistical package written in R that maps each analysis step in our study protocols to a separate small programming task to facilitate implementation.

Discussion

Key challenges included limitations in data access, availability of important confounding variables, small sample sizes, a high proportion of missing values, a lack of consistent time-stamps and irregular follow-up visits. Our findings provide insights into the data, design and analysis aspects crucial for reliable results and contribute to the development of a transparent framework for designing robust non-randomized studies based on registry data.

Simulation study on methods for ordinal data in central statistical monitoring and their application on the German Multiple Sclerosis Registry

15:00-15:15

Presenter: Firas Fneish, David Ellenberger

Authors: Firas Fneish [1]; David Ellenberger [1]; Niklas Frahm [1]; Alexander Stahmann [1]; Gerhard Fortwengel [2]; Frank Schaarschmidt [3]

Affiliations: [1] MS-Register der DMSG, Bundeverband e.V., MS Forschungs- und Projektentwicklungs-gGmbH, Hannover; [2] Hochschule Hannover, Hannover; [3] Leibniz Universität Hannover, Institut für Zellbiologie und Biophysik, Abteilung Biostatistik, Hannover

Introduction

Central statistical monitoring (CSM) is a well-established approach to strengthen data quality in clinical research by using statistical methods to detect irregularities that may indicate issues in trial conduct. Numerous studies have explored statistical tools to support CSM and identify atypical site behaviors or unexpected data patterns. In earlier work, we demonstrated the value of comparing individual centers with the Grand Mean (GM) as a flexible framework applicable across various data types1 and evaluated models implementing such comparisons for binomial, continuous, and time-to-event outcomes. As statistical methodology for ordinal data remains limited, we extend this work by assessing whether similar comparisons can be conducted using cumulative link models and non-parametric ordinal methods implemented in the R packages ordinal, emmeans, and nparcomp. To illustrate practical relevance, we apply this approach to Real-World Data (RWD) from the German Multiple Sclerosis Registry2.

Methods

In a Monte-Carlo simulation study, we investigated whether these methods can control type I error when comparing centers to the GM under varying parameters. Ordinal data were generated using a latent variable approach, where latent data were sampled from the logistic distribution and then categorized by fixed thresholds into 4 or 10 categories. Simulations were run for balanced and unbalanced scenarios covering a range of settings that could be found in clinical trials. For each scenario, nsim = 1000 data sets were generated to compare both methods.

Results

The simulations indicate similar performance of both methods where sample size ni >10 for balanced and unbalanced scenarios. However, for smaller sample sizes ni ≤10 nparcomp is more liberal when controlling the type I error. The approach is implemented on GMSR data to evaluate disease severity of patients after diagnosis. Methods identified centers exhibiting deviations from the GM, highlighting sites that may require further investigation.

Discussion

Both approaches can be applied simultaneously - however - nparcomp tends to be more liberal in controlling the type I error rate compared with the cumulative link model, particularly in situations where centers have very small sample sizes ($n \leq 10$). In addition, the cumulative link model allows the inclusion of covariates, such as center type, when performing the GM comparisons.

References

1. Fneish, F. et al. Application of Statistical Methods for Central Statistical Monitoring and Implementations on the German Multiple Sclerosis Registry. Ther. Innov. Regul. Sci. 57, 1217-1228 (2023).

2. Ohle, L.-M. et al. Chances and challenges of a long-term data repository in multiple sclerosis: 20th birthday of the German MS registry. Sci. Rep. 11, 13340 (2021).

RISK-PRINCIPE: Local Implementation of a Hospital-Onset Bacteremia Detection Algorithm and Proof-of-Concept for a Whole Genome Sequence-based Pathogen Surveillance Module

15:15-15:30

Presenter: Antje Bauer

Authors: Antje Bauer [1]; Seven Johannes Sam Aghdassi [2]; Michael Behnke [2]; Carola Bothe [3]; Luis Alberto Peña Diaz [2]; Martin Kaase [4]; Johannes Oehm [5]; Ferenc Darius Rüther [2]; Franziska Schuler [6]; Alexander Mellmann [1]

Affiliations: [1] Institute of Hygiene, University Hospital Münster, Münster; [2] Institute for Hygiene and Environmental Medicine, Charité - Universitätsmedizin Berlin, Berlin; [3] Department for Infectious Disease Epidemiology, Robert Koch Institute, Berlin; [4] Institute of Hygiene and Infectiology, University Hospital Göttingen, Göttingen; [5] Institute of Medical Informatics, University Münster, Münster; [6] Institute of Medical Microbiology, University Hospital Münster, Münster

Introduction

Hospital-onset bacteremia (HOB) is a key indicator of healthcare-associated infections. Based on the PRAISE HOB algorithm and consensus definitions (Euro Surveill. 202530:2400613) the RISK PRINCIPE consortium developed a modified version of the algorithm which can be applied to routine care data to detect HOBs. To assess potential patient-to-patient transmission of detected pathogens, whole genome sequencing (WGS)-based typing can be applied. Here, we report in a proof-of-concept study the institutional implementation and validation of the modified algorithm alongside the prospective design of a WGS-based surveillance workflow.

Methods

The algorithm was locally implemented by integrating patient movement data and microbiological blood culture results. Validation compared our implementation against the reference algorithm (Charité) using real-world bacteremia cases. Subsequently, the prospective WGS-based workflow for transmission and outbreak detection was designed using core genome multilocus sequence typing (cgMLST) including a validation using real-world bacteremia cases spiked with WGS datasets originating from previous transmission events (J Clin Microbiol. 201452:2365).

Results

The HOB detection algorithm was successfully implemented and validated using The algorithm was successfully implemented and validated using a real-world dataset (2018-2024) of 18,092 patients comprising 107,732 movement records and 84,492 blood cultures, demonstrating full concordance with the reference algorithm. The complete surveillance workflow - encompassing HOB classification, primary/secondary bacteremia stratification, and cgMLST-based transmission detection - was documented in a detailed process flowchart. cgMLST-based transmission detection relies on matching HOB cases against an institutional cgMLST database with temporal stratification for cluster identification. The surveillance workflow was configured to generate automated alerts upon detection of recurrent identical cgMLST types within the institutional database, enabling real-time identification of potential transmissions.

Discussion

The successful implementation and validation of the modified algorithm provide a robust foundation for an integrated, data-driven pathogen surveillance. Integration of automated case classification with molecular analysis facilitates a shift from reactive case identification to proactive transmission detection. This modular framework demonstrates the feasibility of institutional adaptation of RISK PRINCIPE and provides the infrastructure for next-phase integration of real-time cgMLST analysis into hospital surveillance, e.g. automated sequencing triggers based on organism-specific HOB thresholds at the ward level and institution-wide baselines, enabling fully automated real-time transmission detection to prevent larger outbreaks.

Cancer Screening

Wed, 30.09.26, 14:00–15:30, Room: N4

Chair: Tadas Žvirblis

Does Imaging Modality Matter for Biomedical AI? Benchmarking Eight CNN Architectures on Paired Contrast-Enhanced and Standard Mammography Images

14:00-14:15

Presenter: Hanife Avci

Authors: Hanife Avci [1]; Gamze Durhan [2]; Hüseyin Bülüç [3]; Meltem Gülsün Akpinar [2]; Figen DEMİRKAZIK [2]; Jale Karakaya [1]

Affiliations: [1] Department of Biostatistics, Faculty of Medicine, Hacettepe University, Ankara; [2] Department of Radiology, Faculty of Medicine, Hacettepe University, Ankara; [3] Department of Radiology, Mehmet Akif Ersoy State Hospital, Çanakkale

Introduction

Full-Field Digital Mammography (FFDM) is widely used for detecting structural breast changes but has limited sensitivity in dense breast tissue. Contrast-Enhanced Mammography (CEM), particularly Contrast-Enhanced Spectral Mammography (CESM), improves lesion visibility through contrast enhancement and dual-energy imaging. However, its advantage over FFDM in deep learning-based classification remains unclear. This study aims to compare the performance of advanced image processing and deep learning models using paired CESM and FFDM images from the same patients.

Methods

Images from a retrospective dataset of 358 patients (2,754 images 1,377 CESM and 1,377 FFDM) were acquired in DICOM format and processed using a standardized pipeline including VOI LUT correction, intensity normalization, CLAHE enhancement, and Otsu-based breast segmentation. The dataset was split into training (80%), validation (10%), and test (10%) sets. Five-fold cross-validation was performed, with each fold trained for 30 epochs at 384×384 resolution. Hyperparameters were optimized, and AdamW was used throughout. Eight CNN models were trained separately on CESM and FFDM under identical settings for normal/malignant classification. Experiments were implemented in Python 3.12 with PyTorch on an NVIDIA GeForce RTX 5080 GPU. Performance was evaluated using standard classification metrics.

Results

Across all models, CESM consistently outperformed FFDM. Xception achieved the highest AUC on CESM (AUC=0.960, F1=0.88), followed by EfficientNet-B2 (AUC=0.922), DenseNet-121 (AUC=0.893), MobileNet-v2 (AUC=0.879), and ResNet-50 (AUC=0.840). On FFDM, EfficientNet-B0 achieved the

highest AUC (AUC=0.840). LeNet-5 showed the weakest performance in both modalities, confirming the limitations of shallow architectures. Overall, CESM appears more informative than FFDM for deep learning-based breast cancer classification.

Discussion

Although CESM has been used clinically since 2011, its adoption remains lower than FFDM. In this study, CNNs trained on CESM consistently outperformed FFDM across all eight models, with the largest AUC gains in Xception (ΔAUC=+0.206) and DenseNet-121 (ΔAUC=+0.114), supporting CESM for AI-assisted breast cancer screening. Compared with breast MRI, CESM is more accessible and cost-effective while preserving enhancement information. This paired-patient design is among the first to compare CESM and FFDM under identical conditions, reducing patient-level variability and providing a reproducible benchmark for breast cancer detection systems. Inception-v3 training, subgroup analyses, and a Python-based web tool will be completed before the congress presentation.

Impact of Transfer Learning and Model Capacity on Patch-Level Metastasis Detection across Training Data Regimes

14:15-14:30

Presenter: Ellena Spieß

Authors: Ellena Spieß [1, 2]; Reinhold von Schwerin [2]; Friederike Liesche-Starnecker [1]; Daniel Hieber [1]

Affiliations: [1] Institute of Neuropathology, Ulm University Medical Center, Faculty of Medicine, Ulm University, Ulm; [2] Ulm University of Applied Sciences, Ulm

Introduction

The development of deep learning models for histopathology frequently relies on transfer learning (TL) to overcome the scarcity of annotated medical data. However, the efficacy of this approach under varying degrees of data availability is still not fully understood. This study systematically evaluates the impact of TL and model capacity on patch-level breast cancer metastasis detection across different PatchCamelyon training data regimes.

Methods

Three network configurations were evaluated on the PatchCamelyon (PCam) dataset [1]: a frozen ResNet18, a partially fine-tuned ResNet18, and a custom small convolutional neural network (Small-CNN) trained from scratch. Experiments were conducted across three training data regimes using 25%, 50%, and 100% of the 262,144 available training patches.

Results

The results demonstrate that while the partially fine-tuned ResNet18 achieved the highest performance under limited-data conditions (AUPRC 0.9020 at 25% data), the compact SmallCNN trained from scratch outperformed all pre-trained variants when the full dataset was available (AUPRC 0.9279). The frozen ResNet18 consistently yielded the lowest performance across all data regimes.

Discussion

These findings indicate that while TL is highly beneficial for data-scarce scenarios, compact domain-specific architectures can be more effective and computationally efficient given sufficient task-specific data. The results show that the optimal modeling strategy in computational pathology depends critically on the volume of available training data.

[1] Bejnordi BE, Veta M, Van Diest PJ, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA. 2017318(22):2199-2210.

The full paper of this abstract will be published in the GMDS series "Studies in Health Technology and Informatics" with doi: 10.3233/SHTI261004, URL:https://dx.doi.org/10.3233/SHTI261004

Natural history model of prostate cancer using PSA testing data

14:30-14:45

Presenter: Birzhan Akynkozhayev

Authors: Birzhan Akynkozhayev [1]; Benjamin Christoffersen [2]; Anna Lantz [2]; Fredrik Strand [2]; Mark Clements [2]

Affiliations: [1] Karolinska Institutet, Stockholm; [2] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm

Introduction

Prostate-specific antigen (PSA) testing is widely used in Sweden despite the absence of a national population-based screening program. Most testing occurs opportunistically, producing irregular patterns in which only a small proportion of PSA tests lead to a biopsy, there the decision to biopsy may depend on biomarker levels, and diagnostic pathways are complex. Natural history screening models often assume regular screening protocols, which are not well suited to the opportunistic setting. We aimed to develop a computationally efficient framework for reconstructing disease natural history from such real-world data.

Methods

We developed a joint screening model that simultaneously estimates biomarker trajectories conditional on disease onset and disease progression from routinely collected PSA data, while explicitly accounting for biomarker-dependent biopsy probabilities. The model was implemented in C++ with

memory-efficient automatic differentiation for analytical gradients. We conducted an extensive simulation study evaluating the model under different onset distributions, screening schedules, and an opportunistic testing scenario with sparse screen-detected cases. We then fitted the model to data from 2.26 million men in Stockholm, Sweden.

Results

Using simulation data, the model performed well across onset distributions and screening schedules, particularly under Weibull and Beard-Makeham specifications. Despite the challenging opportunistic scenario, with 94.9% censoring and only 0.39% screen-detected cases, the model performed well, with sample sizes of up to 50,000 needed to achieve precision comparable to that of regular screening designs. In the Stockholm cohort, half of men were predicted to have had prostate cancer onset by 80.6 years (95% CI: 80.2-81.0). After onset, half of men were predicted to have a clinical diagnosis within 31.8 years (95% CI: 31.5-32.2). Pre-onset PSA doubling time was 26.9 years, accelerating to 4.6 years after preclinical cancer onset.

Discussion

We developed a natural history model designed for opportunistic testing settings. Applied to Stockholm data, the model suggested that PSA growth associated with preclinical prostate cancer is higher than reported in US-based studies. The framework is computationally efficient, flexible, and modular, accommodating easy changes to disease progression distributions, biopsy probability models, sensitivity assumptions, and biomarker functional form. It is available as an R package, with CRAN submission planned.

Correcting lead-time bias in breast cancer screening: A continuous tumor growth model applied to registry data

14:45-15:00

Presenter: Marius Vachia Robert

Authors: Marius Vachia Robert [1]; Hélène Jacqmin-Gadda [2]; Virginie Rondeau [2]; Marie Poiseuil [2]

Affiliations: [1] National Institute of Health and Medical Research (Inserm U1219, Bordeaux population health), University of Bordeaux, Bordeaux; [2] National Instutite of Health and Medical Research (Inserm U1219, Bordeaux population health), University of Bordeaux, Bordeaux

Introduction

Estimating the survival benefit of breast cancer screening is often compromised by lead time bias, resulting in an artificial overestimation of survival time due to earlier detection rather than delayed death. Lead time refers to the interval between disease detection through screening and the time symptoms would have appeared in the absence of screening. To correct this bias, we developed a statistical methodology for registry data.

Methods

Building upon Abrahamsson et al. (2020), we developed a new continuous tumor growth model and adapted a lead-time bias correction formula. The approach is based on distributions describing tumor growth dynamics, age at tumor onset, time to symptom detection, breast cancer screening sensitivity. We introduced new hypotheses regarding the distribution of age at tumor onset, and implemented these corrections for both individual lead-time estimates and lead-time-adjusted Kaplan-Meier survival curves. To validate this framework, a simulation study was conducted to assess estimator performance, AIC-based model selection, and quantify the overall magnitude of lead-time bias across different screening scenarios. Finally, the methodology was applied to real-world data from a French cancer registry to show its clinical utility.

Results

Simulation results show that model selection via AIC is robust across different biological assumptions. Coverage rates around 95% for both model parameters and adjusted survival probabilities confirm the stability of our estimations. Applied to French registry data, the model provided reliable estimations for tumor growth rate distributions, size at symptomatic detection, screening sensitivity, and age at tumor onset. The lead-time correction revealed significantly different survival curves compared to unadjusted estimates, highlighting the substantial impact of this bias on observed clinical outcomes.

Discussion

The simulation study provides empirical validation of estimator stability, establishing a practical framework for lead-time correction. Application to French registry data demonstrates the feasibility of this approach while considering alternative distributions of age at cancer onset. These results emphasize the necessity of lead-time adjustments for accurate survival analysis and for evaluating the benefits of screening programs for populations.

Abrahamsson et al. (2020). Continuous tumour growth models, lead time estimation and length bias in breast cancer screening studies. Statistical methods in medical research.

Estimating lead time and overdiagnosis in the Slovenian breast cancer screening programme using the MOCCI method

15:00-15:15

Presenter: Bor Vratanar

Authors: Bor Vratanar; Maja Pohar Perme

Affiliations: Institute for Biostatistics and Medical Informatics University of Ljubljana, Ljubljana

Introduction

Cancer screening enables earlier cancer diagnosis and treatment. The interval between screen detection and the time at which the cancer would have been diagnosed clinically in the absence of screening is called the lead time. If a patient's tumour detected by screening would never surface clinically (e.g. because the tumour is non-progressive), the patient is considered overdiagnosed. Estimating these quantities is essential for evaluating the benefits and harms of cancer screening programmes. In this talk, we introduce MOCCI (Minimizing Observed and Counterfactual Cancer Incidence), a novel causal-inference method that we developed to estimate the distributions of lead time and overdiagnosis.

Methods

MOCCI compares age- and calendar-stratified cancer incidence between individuals invited to screening and those not invited and estimates the lead time distribution that minimizes the incidence difference between the two groups. The probability of overdiagnosis is then derived by comparing the model-predicted lead time with the time to death from causes other than the target cancer. We applied MOCCI to the Slovenian breast cancer screening programme. Although the programme was introduced in 2008, it initially operated only in the central region and was gradually expanded to all regions by 2019. This phased rollout created a natural experiment in which some women were invited to screening while others were not, providing the data required for MOCCI.

Results

As expected, cancer incidence in the invited cohort was much higher (due to overdiagnosis) and shifted towards younger ages than in the non-invited cohort. Applying MOCCI to these data, we estimated that 33% (95% CI: 25%-42%) of screen-detected cancers were overdiagnosed and that the mean lead time among progressive cancers was 1.7 years (95% CI: 1.3-2.9).

Discussion

Our results are based on a non-randomized comparison of invited and non-invited populations although we undertook several steps to make these groups as comparable as possible, comparability remains an untestable assumption. Note also, that estimates of both lead time and overdiagnosis are country-specific: they depend on screening frequency, eligibility criteria, mammography sensitivity and other programme characteristics, and may not generalise directly to other breast cancer screening programmes.

Evidence-Informed Guideline Development for Cervical Cancer Screening in Austria - A Modelling Study Commissioned by the Austrian Federal Ministry of Health to Advise the National Committee for Cancer Screening

15:15-15:30

Presenter: Gaby Sroczynski

Authors: Gaby Sroczynski [1]; Lára R. Hallsson [2]; Lennart Koch [2]; Milena Mühler [2]; Beate Jahn [2]; Nikolai Mühlberger [2]; Andreas Widschwendter [3]; Elmar Joura [4]; Nadja Taumberger [5]; Olaf Reich [6]; Peter Hillemanns [7]; Wegene Borena [8]; Monika Hackl [9]; Jane J. Kim [10]; Barbara Fröschl [11]; Christina Dietscher [12]; Eva Schernhammer [13, 14, 15]; Gerald Gartlehner [16, 17]; Uwe Siebert [2, 10, 18, 19, 20]

Affiliations: [1] Department of Public Health, Health Services Research, and Health Technology Assessment, UMIT TIROL - University for Health Sciences and Technology, Hall in Tirol; [2] Department of Public Health, Health Services Research, and Health Technology Assessment, UMIT TIROL - University for Health Sciences and Technology, Hall in Tirol; [3] Department of Obstetrics and Gynaecology, Medical University Innsbruck, Innsbruck; [4] Department of Obstetrics and Gynaecology, Medical University Vienna, Vienna; [5] Department of Obstetrics and Gynaecology, Medical University Graz, Hall in Tirol; [6] Department of Obstetrics and Gynaecology, Medical University Graz, Graz; [7] Department of Gynaecology and Obstetrics, Hannover Medical School, Hannover; [8] Institute of Virology, Department of Hygiene, Microbiology, Medical University of Innsbruck, Innsbruck; [9] Statistics Austria, Austrian National Cancer Registry, Vienna; [10] Center for Health Decision Science, Department of Health Policy & Management, Harvard T.H. Chan School of Public Health, Boston; [11] Gesundheit Österreich GmbH, Vienna; [12] Austrian Ministry of Labour, Social Affairs, Health, Care and Consumer Protection, Vienna; [13] Center for Public Health Department of Epidemiology, Medical University of Vienna, Vienna; [14] Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston; [15] Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston; [16] Department for Evidence-based Medicine and Evaluation, University for Continuing Education Krems, Krems; [17] Center for Public Health Methods, RTI International, Research Triangle Park, North Carolina; [18] Division of Health Technology Assessment and Bioinformatics, ONCOTYROL-Center for Personalized Cancer Medicine, Innsbruck; [19] Institute for Technology Assessment and Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston; [20] Center for Health Decision Science, Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston

Introduction

In Austria, opportunistic cervical cancer screening still includes annual cytology for women as of age 18. Actual European guidelines recommend organized screening with five-yearly HPV-testing as of age 30.1 This decision-analytical modelling study was commissioned by the Austrian Federal Ministry of Health (BMASGPK) to advise the Austrian National Committee for Cancer Screening (ANCCS) in optimizing the national screening policy. We evaluated the long-term benefit-harm trade-offs of different cervical cancer screening strategies.

Methods

We established an Expert Panel consisting of members of the ANCCS, national and international clinical and modelling experts. In several consensus workshops with a modified Delphi procedure, the Expert Panel discussed and agreed on the study design, framework, strategies, and primary study endpoints. Using a validated Markov state-transition model, we evaluated 52 screening strategies differing by primary test (Pap, HPV testing, or co-testing), screening interval (1, 3, 5 years), and age

limits (lower age at start 21, 25, upper age for end 75 or no limit). Austrian epidemiological and clinical data along with international test performance data from meta-analyses were used. Primary outcomes included life-years gained (LYG), additional positive screening findings (APSF) as a proxy for potential psychological and physical harms, and the stepwise incremental harm-benefit ratio (IHBR). Deterministic multiway sensitivity analyses were performed to assess robustness of the base-case results.

Results

Both benefits and harms increased with greater screening intensity such as wider age ranges and shorter screening intervals. Among non-dominated strategies, triennial Pap from ages 30-75 or 25-75, followed by 21-75 achieved the lowest IHBRs with 5, 5 and 16 APSF/LYG vs. no screening, respectively. Transitioning to 5-yearly HPV testing at ages 30-75 further improved benefits, with IHBRs of 20 or 38 APSF/LYG when combined with triennial or annual Pap at ages 21-29. Triennial HPV testing at ages 30-75 yielded 81 APSF/LYG, while removing the upper age limit increased the IHBR to 569 APSF/LYG. Co-testing every three years from age 30 produced the highest IHBR of 2,099 APSF/LYG. Results were robust in sensitivity analyses for most variations.

Discussion

Our findings provide an evidence base to inform national decision-making on cervical cancer screening in Austria.

Time-to-Event Methods and Meta-Analysis

Wed, 30.09.26, 14:00–15:30, Room: N5

Chair: Theodoros Evrenoglou

Testing for subgroup treatment effect consistency in the Cox model

14:00-14:15

Presenter: Lukas Koletzko

Authors: Lukas Koletzko [1]; Holger Dette [1]; Björn Bornkamp [2]; Frank Bretz [2]; Zoe Lange [1]

Affiliations: [1] Ruhr-Universität Bochum, Bochum; [2] Novartis Pharma AG, Basel

Introduction

If the overall result of a Phase III trial is positive, an important next step is to assess whether treatment effects are consistent across clinically relevant subgroups in order to strengthen the applicability of the drug and justify broad use in heterogeneous patient populations. For time-to-event endpoints, however, formal testing methodology for confirming subgroup consistency does not exist. Our paper addresses this gap for the Cox proportional hazards model by formulating subgroup consistency as an equivalence problem rather than a classical interaction-testing problem, so that rejecting the null hypothesis allows to conclude that subgroup differences are medically irrelevant.

Methods

We consider a Phase III setting with two complementary subgroups and investigate two consistency questions: first, whether the subgroup-specific treatment effects are consistent with each other and second, whether the subgroup effects are consistent with the overall population effect. Under the Cox model, both problems reduce to equivalence hypotheses involving only the (weighted) treatment-by-subgroup interaction parameter. Based on this reduction, we develop classical TOST-type procedures as well as new bootstrap tests in order to test for consistency. We establish the asymptotic validity of the proposed methods and also include a simulation study. The methodology is exemplarily applied to a case study which is motivated by the CANTOS cardiovascular trial.

Results

Overall, the paper provides a formal equivalence-testing framework for testing subgroup treatment effect consistency with survival outcomes in the Cox proportional hazards model for two subgroups. Our simulation results indicate that the proposed bootstrap tests are consistently more powerful than classical TOST-type tests, particularly for small sample sizes and higher levels of censoring.

Discussion

For time-to-event data, testing subgroup treatment effect consistency yields an inherently non-parametric problem, since treatment effects are naturally time-varying functions. In our work, we reduced the functional consistency hypotheses to hypotheses involving only a one- or two-dimensional parameter under the proportional hazards model, however we note that such a parametric reduction of the problem may not work in more general survival models. Consequently, parametric bootstrap tests such as those proposed in this work will not be available and the consistency problems have to be treated with different methods.

Optimizing an early-late contrast within a single survival curve using window-specific mean survival

14:15-14:30

Presenter/Author: Lubomir Stepanek

Affiliations: First Faculty of Medicine, Charles University, Prague; Faculty of Informatics and Statistics, Prague University of Economics and Business, Prague; Institute of Computer Science, Czech Academy of Sciences, Prague

Introduction

Survival analysis usually focuses on comparisons between groups, but important information may also lie within a single survival curve. We study whether the early and late phases of one group's survival experience differ systematically, and whether this difference can be summarized and optimized. For a candidate split time c, we compare the early interval (0, c) and the late interval (c,tau), where tau is a fixed upper time horizon. Small c yields more distinct early-late summaries but uses less follow-up information, while larger c improves precision but makes the two summaries increasingly similar. This creates a natural optimization problem.

Methods

For one Kaplan-Meier curve hat S(t), we defined the early and late window-specific mean survival as

ME(c)=frac1cint0c hat S(t),dt,qquad ML(c)=frac1tau-cintc^tau hat S(t),dt.

Their difference,

D(c)=ME(c)-ML(c),

was taken as the raw early-late contrast. To compare split times fairly, we standardized this contrast using the covariance structure of the Kaplan-Meier process with Greenwood-type plug-in estimation,

Z(c)=fracD(c)sqrtVarD(c).

We studied the population-level optimizer c^*=argmaxc Z(c) under selected parametric survival models and linked c^* to an expected event index for interpretation. Models included a piecewise exponential model with one hazard change-point, log-logistic and lognormal models with non-monotone hazards, and a two-component exponential mixture representing latent heterogeneity. Exponential and Weibull models were examined as monotone-hazard benchmarks.

Results

For models with structural time variation, the optimal split corresponded to an interpretable model feature: the hazard change-point in the piecewise exponential model, the hazard peak in the log-logistic and lognormal models, and the transition between dominant mixture components in the exponential-mixture model. In contrast, for exponential and Weibull models, Z(c) changed monotonically with the split, indicating that no meaningful internal optimum exists.

Discussion

The proposed early-late contrast provides an exploratory within-curve diagnostic for detecting temporal structure, latent heterogeneity, or phase transitions in survival dynamics. By optimizing over split times within a single survival curve, the method extends survival analysis from between-group comparison to within-curve characterization.

Restricted mean time lost ratio regression: Percentage of restricted mean time lost due to specific cause

14:30-14:45

Presenter/Author: Thomas Scheike

Affiliations: Section of Biostatistics, Department of Public Health, University of Copenhagen, Copenhagen

Introduction

Restricted mean survival time (RMST) has gained increasing attention as an interpretable summary measure of survival, representing the average time alive up to a prespecified time horizon. In competing risks settings, RMST can be decomposed into cause-specific components through the restricted mean time lost (RMTL), allowing attribution of time lost to distinct causes of death. While regression approaches for RMTL have been proposed, interpretation of covariate effects remains challenging when multiple competing events contribute to overall time lost. In applied research, it is important to understand not only the total RMTL but also how covariates influence the distribution of time lost across different causes. We propose a regression framework that directly models the relationship between total RMTL and cause-specific RMTL components.

Methods

We introduce RMTL ratio regression models in which the proportion of total RMTL attributable to specific causes is modeled through regression techniques. Logistic regression formulations provide odds ratio interpretations of covariate effects on the percentage of time lost due to a given cause. Alternatively, log-odds product models offer direct estimation of percentages for key prognostic factors. The framework naturally extends to the full decomposition of RMTL across multiple competing events using multinomial regression or extensions of log-odds product models. Estimation is performed using inverse probability weighting techniques to account for censoring and ensure consistent estimation.

Results

The proposed modeling framework provides interpretable measures of how covariates influence the relative contribution of competing causes to total time lost. Logistic formulations yield straightforward odds ratio interpretations, while log-odds product models facilitate direct estimation of clinically meaningful percentages. The inverse probability weighting approach performs well in practical applications and supports stable estimation across a range of competing risk scenarios.

Discussion

RMTL ratio regression offers a flexible and interpretable approach to analyzing competing risks data. By focusing on the proportional contribution of each cause to total restricted mean time lost, the framework enhances understanding of covariate effects beyond traditional survival or cause-specific analyses. The proposed methodology provides a practical tool for applied researchers seeking clinically interpretable measures in competing risks settings and can be readily extended to multiple causes of death.

A Poisson approximation for two-stage meta-analysis of time-to-event data

14:45-15:00

Presenter: Ian White

Authors: Ian White [1]; Becky Turner [1]; Richard Riley [2]; David Fisher [1]

Affiliations: [1] UCL Innovative Clinical Trial Unit, London; [2] University of Birmingham, Birmingham

Introduction

Two-stage meta-analysis of randomised trials typically relies on a Normal approximation to the study-specific log-likelihoods, but this does not work well in all settings. Focussing on time-to-event outcomes, we aimed (1) to explain why the two-stage Normal method can yield biased treatment effect estimates for time-to-event outcomes, (2) to quantify the magnitude and direction of bias under realistic conditions, and (3) to develop and evaluate a novel method to reduce bias.

Methods

We use profile likelihood to show how and when the Normal-approximation method causes bias. We propose a new Poisson-approximation method which additionally uses event counts per group to better approximate the correct likelihood (i.e. the likelihood that would be obtained from individual participant data). We compare the two methods with each other and with one-stage Cox analysis (reference method) in a simulation study in which event incidence, treatment effect heterogeneity, numbers of studies and allocation ratios were varied.

Results

Profile likelihood shows that departure from Normality is mainly driven by the ratio of the event counts in the trial arms. When all studies have roughly equal group sizes, the two-stage Normal approximation is biased towards the null. However, when the group sizes are unequal but in approximately the same ratio in all studies, the bias is towards the log ratio of group sizes.

Simulation confirms these findings. It also shows that treatment effect estimates from the two-stage Normal approximation can deviate markedly from the reference method, while those from the two-stage Poisson approximation do not. The Poisson approximation reduces bias in two-stage analysis compared with the Normal approximation in data with few events, and gives similar results to the Normal approximation in data with many events. However, in random-effects models fitted to heterogeneous data, the Poisson approximation gave somewhat worse under-coverage than the Normal approximation.

Discussion

The Poisson approximation method appears to have lower bias than the Normal approximation method. However, further refinement of the method and more extensive simulations with random-effects models are needed before it can be broadly recommended.

Beyond Hazard Rate: Multivariate Restricted Mean Survival Time Meta-Analysis for Single-Arm studies

15:00-15:15

Presenter: Georgios Seitidis

Authors: Georgios Seitidis [1]; Ruben Van Eijk [2]; Stavros Nikolakopoulos [1]

Affiliations: [1] Department of Psychology, University of Ioannina, Ioannina, Greece, Ioannina; [2] Department of Neurology, UMC Utrecht Brain Center, University Medical Center Utrecht, Heidelberglaan 100, 3584 CX, Utrecht, The Netherlands, Ioannina

Introduction

When synthesizing survival outcomes from single-arm studies, Restricted Mean Survival Time (RMST) provides a practical and robust summary of survival on the time scale. However, because follow-up varies across studies, RMST may not be available at the same truncation time for all studies. The common practice is to perform separate univariate meta-analyses at each truncation time, but this progressively excludes shorter follow-up studies at later time points, yielding to imprecise and non-reliable estimates.

Methods

We propose a Bayesian multivariate meta-analysis framework where RMST estimates at truncation times are modeled jointly as correlated outcomes and as a function of truncation time. The model accommodates cases where RMST is not available for all truncation times by properly adjusting the design matrix, allowing studies with shorter follow-up to still contribute instead of being excluded. Dependence across time points is captured using parsimonious autoregressive covariance structures, reducing the number of parameters to be estimated. We extend the framework to meta-regression setting to assess whether RMST differ by study-level characteristics and whether such differences vary with follow-up. Performance is evaluated through simulations and a meta-analysis of Amyotrophic Lateral Sclerosis (ALS) trials derived from the EVIDENCE project.

Results

Across simulation scenarios with varying number of studies and missing rates across time points, the proposed multivariate framework produced less biased and more precise estimates than univariate meta-analyses. Improvements were observed across truncation times, with the most evident differences at later time points, where univariate meta-analyses are driven by fewer contributing studies, producing unstable and in some cases implausible estimates. In a meta-analysis of ALS trials using 10 studies and 2783 patients, we applied the multivariate framework to assess whether administration route affects survival. Multivariate meta-regression provided more precise and stable estimates, supporting no differences between administration routes.

Discussion

Multivariate modeling of RMST across truncation times offers an efficient and solid approach to synthesize single-arm studies with different follow-up periods. By borrowing strength across time points, the proposed method could provide more powerful and plausible estimates. Furthermore, it can be used in a network meta-analysis setting to inform whether nodes can be combined, thereby helping to connect an otherwise disconnected network.

Network meta-analysis (NMA) of prediction model evaluations using individual participant data (IPD) and large clustered data sets: guidance for their conduct

15:15-15:30

Presenter: Maerziya Yusufujiang

Authors: Maerziya Yusufujiang [1]; Ewoud Schuit [1]; Richard Riley [2]; Ewout Steyerberg [1]; Karel Moons [1]; Valentijn de Jong [1]

Affiliations: [1] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht; [2] University of Birmingham, Birmingham

Introduction

Prediction models for the same clinical task are often evaluated across different settings. Because not all models are validated in every dataset, direct head-to-head comparisons are often limited, although indirect comparisons may still be possible. Network meta-analysis (NMA) provides a framework for combining direct and indirect evidence across multiple datasets however, guidance on synthesising prediction model performance with NMA, particularly with individual participant data (IPD), remains limited.

Objective: To provide guidance for direct and indirect comparisons of prediction model performance across studies using NMA in situations where IPD of the studies is available, with attention to core NMA assumptions.

Methods

We provide a step-by-step framework for NMA of prediction model performance using IPD. We illustrate the framework in a case study using the eICU Collaborative Research Database by evaluating six models for in-hospital mortality in patients admitted to the intensive care unit, utilising the IPD from 55 hospitals. To mimic the situation of indirect prediction model comparisons, we also apply our NMA framework by (randomly) removing ~25% and ~50% of hospital-specific model performance estimates.

Results

The framework demonstrated how NMA can be used to compare prediction model performance across a connected network of direct and indirect evidence, while explicitly addressing key assumptions, including transitivity, heterogeneity, and consistency. In the case study, the model-comparison network remained connected even after reducing data availability, although it contained fewer hospital-specific model comparisons. Between-hospital heterogeneity was observed, while no major inconsistency was identified. As a result, performance estimates and ranking estimates became less precise. Overall, comparative performance patterns were broadly similar across scenarios, but uncertainty overlapped among the top-performing models as fewer hospital-specific evaluations were available.

Discussion

NMA methods can be applied to synthesize and compare the performance of prediction models in the presence of IPD, allowing simultaneous comparison of multiple models using both direct and

indirect evidence, even with different numbers of prediction models evaluated across clusters. However, between-dataset heterogeneity and reduced data availability increase uncertainty in comparative estimates and rankings. Rankings should therefore be interpreted alongside effect sizes and uncertainty intervals.

Poster Sessions

Mon 28.09.2026

P1 - Digital Health and Data Systems

Poster area: P1

[28-P1-01] From (Data) Standards to Interoperability: Empirical Insights into Good Health Data Modeling

Presenter: Jendrik Richter

Authors: Lena Elgert [1]; Jendrik Richter [2]; Jannik Fleßner [3]; Darian Liehr [4]; Matthias Katzensteiner [4]; Mona Pape [1]; Nadine Schulz [1]; Dagmar Krefting [5]; Frauke Koppelin [6]; Oliver J. Bott [4]; Klaus-Hendrik Wolf [7]

Affiliations: [1] Peter L. Reichertz Institut für Medizinische Informatik der TU Braunschweig und der Medizinischen Hochschule Hannover, Hannover; [2] University Medical Center Göttingen, Department of Medical Informatics, Göttingen, Göttingen; [3] Jade Hochschule Wilhelmshaven, Oldenburg, Elsfleth, Elsfleth; [4] Hochschule Hannover, Hannover; [5] University Medical Center Göttingen, Göttingen; [6] Jade Hochschule, Oldenburg; [7] Medizinische Hochschule Hannover, Hannover

Introduction

Health data modeling geared towards comprehensive interoperability is crucial for efficient processes and data (re)usability in research and healthcare. This requires the effective planning and implementation throughout the entire data lifecycle - from modeling and analysis to reuse - using appropriate health data standards. [1-3] Against this background, the aim of this study is to inform project planners and data stewards [4] about the selection and use of health data standards.

Methods

With an online survey and an expert workshop insights on the structuring and key aspects of successful health data modeling were gained and discussed. The data were analyzed using descriptive statistics and qualitative content analysis. A process diagram outlines the resulting aspects of health data modeling.

Results

Twenty-six experts in health data modeling completed the online questionnaire. The results cover aspects of modeling medical information models and health data, modeling practice, available aids, and stakeholders involved. A level model derived from the results illustrates the modeling of health data, embedded into the influence levels of regulation, governance, and modeling practice.

Discussion

The classification of examples and best practices highlights the extensive knowledge required for the use and planning of health data standards, as well as the need to consider cross-project objectives such as the harmonization of health data models. Beyond that, a blueprint for modeling health data could and should not be derived since use cases, data sources, data quality and established system architectures are highly heterogeneous. The strategic planning of interoperable data models needs comprehensive knowledge especially regarding further reaching goals like harmonization of models. Available aids e.g. frameworks and guidelines like [5] can support. The level model and further results are intended to support both project planners and data stewards in the selection and use of health data standards, as well as in the organization of modeling activities.

References

[1] Ammon D, Kurscheidt M, Buckow K, Kirsten T, Löbe M, Meineke F et al. Arbeitsgruppe Interoperabilität: Kerndatensatz und Informationssysteme für Integration und Austausch von Daten in der Medizininformatik-Initiative. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2024 67(6):656-67. doi: 10.1007/s00103-024-03888-4. [2] Waltemath D, Beyan O, Crameri K, Dedié A, Gierend K, Gröber P et al. FAIRe Gesundheitsdaten im nationalen und internationalen Datenraum. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2024 67(6):710-20. doi: 10.1007/s00103-024-03884-8. [3] Seixas-Lopes FA, Lopes C, Marques M, Agostinho C, Jardim-Goncalves R. Musculoskeletal Disorder (MSD) Health Data Collection, Personalized Management and Exchange Using Fast Healthcare Interoperability Resources (FHIR). Sensors (Basel) 2024 24(16). doi: 10.3390/s24165175 [4] Katzensteiner, M., Müller, A., GMDS-Arbeitsgruppe 'Datenmanagement in klinischen und epidemiologischen Studien', GMDS Arbeitsgruppe 'Curricula der Medizinischen Informatik'. Data-Steward(-ship) in der Digitalisierung des Gesundheitswesens. Forum der Medizin_Dokumentation und Medizin_Informatik 2022 [cited 2026 Mar 12] (4). Available from: https://www.gmds.de/fmin/user_upload/AG_Datenmanagement/mdi_4_2022_web.pdf. [5] eHealth Network. Refinement of the eHealth European Interoperability Framework (ReEIF). 7th Meeting of the eHealth Network. April 2015 2015 [cited 2026 Jun 23]. Available from: https://health.ec.europa.eu/document/download/7cc56460-46ea-4fd7-9064-ac1a14a2a14e_en

The full paper pertaining to this abstract will be published in the GMDS journal "Medizinische Informatik, Biometrie und Epidemiologie (MIBE)."

[28-P1-02] Beyond Disease Boundaries: Harmonized Core Dataset (CDS) for Clinical Studies of the German Centers for Health Research (DZG)

Presenter: Angela Dedié

Authors: Angela Dedié [1]; Ivona Anastasova [1]; Meike T. Funderich [2]; Jori Kern [3]; Raphael W. Majeed [2]; Katharina Wendlandt [4]; Steffen Cordes [5]

Affiliations: [1] German Center for Diabetes Research e.V. (DZD), Neuherberg; [2] Department of Internal Medicine, Universities of Giessen and Marburg Lung Center (UGMLC), Member of the German Center for Lung Research (DZL), Giessen; [3] Deutsche Konsortium für Translationale Krebsforschung (DKTK), Heidelberg; [4] Department of Internal Medicine, Universities of Giessen and Marburg Lung Center (UGMLC), Member of the German Center for Lung Research (DZL), Giessen (Germany), Giessen; [5] Deutsches Zentrum für Herz-Kreislauf-Forschung e.V. (DZHK), Berlin

Introduction

The German Centers for Health Research (DZG) are dedicated to investigating major diseases that are highly prevalent in the population and often interrelated like diabetes, infectious diseases, mental disorders, cancer, cardiovascular, lung, pediatric and neurodegenerative diseases. Comorbidities and therapy-related adverse effects necessitate a holistic understanding of disease interconnections. To address this need, the DZG have developed a harmonized core dataset (CDS) for clinical studies. This dataset enables standardized and structured collection of clinical data and facilitates analyses across disease boundaries.

Methods

The mandatory collection of a cross-DZG CDS in DZG clinical studies was decided by the DZG Board in 2023. Its parameters were developed and harmonized in a multi-stage consensus process in close coordination with clinicians from all DZG. The resulting CDS comprises 34 mandatory items in the item groups "General Information" and "Diseases" that form the basis for cross-DZG identification and analysis of patient cohorts. In addition, an optional dataset including 100 further items was consented, covering variables relevant to cross-indication scientific research. All items were semantically annotated using the international terminologies LOINC and SNOMED CT to ensure interoperability and facilitate standardized data exchange.

Results

In September 2023, the mandatory DZG CDS was published under an open license and assigned the DOI 10.21961/mdm:46209, to enhance its findability, accessibility and potential for reuse. A standardized update process and a practical guidance document for data collection further support harmonization and interoperability. The DZG CDS supports researchers in identifying patient cohorts across the various research centers and creates new opportunities for collaboration. The extended dataset further enables the prospectively harmonized collection of items in cross-indication studies, thereby facilitating novel insights into disease mechanisms that transcend individual disorders.

Discussion

Since late 2024, the mandatory parameters of the DZG CDS have been collected across all DZG. However, it remains challenging to determine in how many DZG studies the CDS is currently being implemented and for how many study participants the required parameters are available. To address this gap, a DZG working group is currently set up to develop methodologies to define, capture, and report key performance indicators of its implementation.

[28-P1-03] Developing a Cross-Research-Area Basic Data Set: A Coordinated Approach to Metadata Standardization and Harmonization within DZKJ

Presenter: Polina Soloveva, Stephanie Brandt-Heunemann

Authors: Polina Soloveva [1, 2]; Stephanie Brandt-Heunemann [3, 2]; Ariane Ketabi [4, 5]; Robert Kossen [4]; Sabine Hanß [4, 5]; Franziska Radicke [6, 7]; Mandy Vogel [8, 9]; Manfred Reichert [10, 2]; Hans Armin Kestler [1, 11, 2]

Affiliations: [1] Institute of Medical Systems Biology, Ulm University, Ulm; [2] German Center for Child and Adolescent Health (DZKJ), partner site Ulm, Ulm; [3] Department of Paediatrics and Adolescent Medicine, Division of Paediatric Endocrinology and Diabetes, Ulm Universiy Medical Center, Ulm; [4] Department of Medical Informatics, University Medical Center Göttingen (UMG), Göttingen; [5] German Center for Child and Adolescent Health (DZKJ), partner site Göttingen, Göttingen; [6] Section Healthcare Epidemiology and Community Health, Institute for Community Medicine, University Medicine Greifswald, Greifswald; [7] German Center for Child and Adolescent Health (DZKJ), partner site Greifswald, Greifswald; [8] University Hospital for Children and Adolescents, Center for Pediatric Research Leipzig, University of Leipzig, Leipzig; [9] German Center for Child and Adolescent Health (DZKJ), partner site Leipzig/Dresden, Leipzig; [10] Institute of Databases and Information Systems, Ulm University, Ulm; [11] Head of Medical Data Integration Center Ulm (MeDIC-Ulm), University Medicine Ulm, Ulm

Introduction

The German Center for Child and Adolescent Health (DZKJ) is a multidisciplinary research initiative integrating seven diverse research areas (RAs). This diversity generates large volumes of heterogeneous data, enabling high-quality cross-domain research while introducing fragmentation in data structures, variable definitions, and data collection methods.

To address this, a common Basic DZKJ Data Set is being developed as a cross-domain standard. In parallel, RA-specific data sets are planned to capture variables relevant to individual research areas. This establishes a shared data layer across RAs, enabling interoperable data use, efficient reuse, rapid study initiation, and advanced analytical approaches such as federated learning.

Methods

A structured, multi-stage process for developing the Basic DZKJ Data Set has been established within the Collaborative Research Data Management Platform of the DZKJ. The approach combines decentralized variable nomination by RAs with centralized review and harmonization, supported by standardized metadata requirements (definitions, data types, units, and collection methods).

Guidelines for metadata specification were formalized and shared with RAs to ensure consistent variable definitions. Each RA proposes a limited set of variables relevant across domains, which are reviewed, compared, and harmonized through a consensus-based process.

Results

A cross-RA nomination process has produced an initial pool of candidate variables for the Basic DZKJ Data Set. This enables the identification of overlaps and inconsistencies across RAs in variable definitions, units, and data collection methods. Building on this process, a guideline for metadata standardization and data harmonization is being developed to support the integration of heterogeneous data sources.

Discussion

The presented approach establishes a structured methodology for deriving a consensus-based variable set that forms the basis for a compact, interoperable dataset. This improves the comparability of research data in child and adolescent health across studies and institutions. Introducing this structure facilitates cross-RA analyses while reducing the effort required for post hoc harmonization.

Although the dataset is currently under refinement, the approach provides a scalable foundation for collaborative research and federated analyses, with the potential to enhance statistical power and robustness.

[28-P1-04] From Data to Discovery: The Collaborative Research Datamanagement Platform for the German Center for Child and Adolescent Health

Presenter: Sabine Hanß, Ariane Ketabi

Authors: Ariane Ketabi [1, 2]; Stephanie Brandt-Heunemann [3, 4]; Felicia Hähner [5, 6]; Sabine Hanß [1, 2]; Robert Kossen [1]; Mathias Manzke [7, 8]; Polina Soloveva [9, 4]; Frank Ückert [5, 6]; Dagmar Krefting [1, 2, 10]

Affiliations: [1] Department of Medical Informatics, University Medical Center Göttingen (UMG), Göttingen; [2] German Center for Child and Adolescent Health (DZKJ), partner site Göttingen, Göttingen; [3] Department of Paediatrics and Adolescent Medicine, Division of Paediatric Endocrinology and Diabetes, Ulm Universiy Medical Center, Ulm; [4] German Center for Child and Adolescent Health (DZKJ), partner site Ulm, Ulm; [5] Institute for Applied Medical Informatics, University Medical Center Hamburg-Eppendorf, Hamburg, Hamburg; [6] German Center for Child and Adolescent Health (DZKJ), partner site Hamburg, Hamburg; [7] Institute and Policlinic of Radiology, Pediatric Radiology and Neuroradiology, University Medical Center Rostock (UMR), Rostock; [8] German Center for Child and Adolescent Health (DZKJ), partner site Greifswald/Rostock, Rostock; [9] Institute of Medical Systems Biology, Ulm University, Ulm; [10] University of Göttingen, Campus-Institute Data Science (CIDAS), Göttingen

Introduction

Current research in pediatric health is hindered by fragmented data management, with limited standardization and harmonization across institutions. This complicates collaboration, reduces reusability, and impedes efficient analysis. The Collaborative Research Data Platform (CRDP) combines decentral and central components and services to optimize usability, data protection and maintainability for the German Center for Child and Adolescent Health (DZKJ). The platform is designed to enable

integrated, FAIR-compliant data management across institutions. It aims to foster FAIR data principles through standardized practices, a common Data Management Plan, guidance for the development of data protection concepts and a central knowledge base. A hybrid architecture balances data security and accessibility, supporting harmonized data management. As the DZKJ expands, the platform will strengthen data integration and collaborative research.

Methods

The CRDP integrates decentralized data sources via standardized interfaces, enabling harmonized management, controlled access, and interoperability. It builds on existing systems to ensure sustainability and avoid redundant development of infrastructures. By supporting research projects from the DZKJ's Research Areas (RAs), the CRDP uses these use cases to develop and refine standardized data management processes for data collection, curation, and FAIR compliance. These processes are then generalized and applied across future initiatives, ensuring consistency and scalability across the DZKJ.

Results

So far, the CRDP has delivered a consortium-wide Data Management Plan, defining responsibilities, FAIR requirements, and data management governance across the DZKJ. A common process for defining the Basic DZKJ Data Set and RA-specific Data Sets, supported by standardized guidelines to ensure consistency and alignment across research areas, has been established. To enhance research coordination, a cohort inventory is being developed, initially in a local data hub and later integrated into the nfdi4health central study hub, allowing researchers to discover existing studies for secondary use of study data.

Discussion

The CRDP lowers technical barriers to data sharing by establishing standardized processes and harmonized data definitions. It supports the discovery, integration, and reuse of data across projects, advancing findability, interoperability, and reusability. These outputs provide a stable, standardized foundation for data management across the DZKJ, built on established national infrastructures and supporting consistent, efficient research practices, and enhancing collaboration.

[28-P1-05] SwissPedHealth National Data Stream Infrastructure: Establishing a National Pediatric Data Framework

Presenter: Xeni Deligianni

Authors: Xeni Deligianni [1]; Tatjana Welzel [1]; Nicole Göbel [2]; Andrea Agostini [3]; Fabiën Belle [4]; Rebeca Mozun [5]; Claudia E. Kuehni [4]; Luregn J. Schlapbach [5, 6]; Eric Giannoni [7]; Klara Posfay-Barbe [8]; Julia Bielicki on behalf of SwissPedHealth consortium [1]

Affiliations: [1] Paediatric Research Centre, University Children's Hospital Basel, University of Basel, Basel; [2] University Hospital of Basel, Basel; [3] Department of Computer Science, ETH Zurich, Switzerland, Zurich; [4] Child and Adolescent Health, Pediatric Routine Data, Institute of Social and Preventive Medicine, University of Bern, Bern; [5] Department of Intensive Care and Neonatology, University Children's Hospital Zurich and Children's Research Centre, University of Zurich, Zurich; [6] Child Health Research Centre, University of Queensland, Brisbane, QLD, Brisbane; [7] Clinic of Neonatology, Department Woman-Mother-Child, Lausanne University Hospital and University of Lausanne, Lausanne; [8] Department of Pediatrics, Gynecology & Obstetrics, Pediatric Infectious Disease Unit, Geneva University Hospitals and Faculty of Medicine, Geneva

Introduction

Pediatric health data in Switzerland remain fragmented across institutions, limiting their reuse potential. SwissPedHealth (SPH)1 is a national pediatric data stream within the Swiss Personalized Health Network (SPHN), designed to enable interoperable and ethically compliant reuse of pediatric health data. It builds on a standardized hospital pediatric data model (SwissPedData) and leverages the SwissPedNet research network.

Methods

Participating hospitals (Basel, Bern, Geneva, Lausanne, Zurich, Lucerne, St. Gallen), prepared their data according to SPHN concepts using the SPHN RDF schema and format, complemented by an SPH-developed schema capturing highly specialized concepts, such as for lung function or multi-omics.

A key regulatory feature of SPH is its central open-ended registry regulation approved by the Ethics Committee Northwest and Central Switzerland (AO-20220018). This enables unbiased use of registry data for quality assessment while research subprojects receive data filtered by the General Consent only after project-specific ethics approvals and data-sharing agreements.

Data flow begins at the data providers, passes through the registry for storage, control and selection and is forwarded to projects based on the respective agreements and ethic approvals.

Dedicated pipelines were developed for control of the data (e.g., concepts' selection) and optional anonymization of initially pseudonymized data.

Results

By the end of 2025, the SPH registry had collected data of 760.861 children and adolescents standardized to international terminologies. All subprojects received the requested health-related data for analysis. However, variations in ethical committees decisions and requirements, mismatches between the RDF format and model and local hospital systems created challenges and delayed timelines and feedback loops.

Recent experience highlighted the importance of early and regular data checks and close collaboration between data providers and the researchers, to address standardization and data quality issues.

Discussion

The SPH registry aggregates standardized clinical pediatrics routine and multi-omics data across Swiss hospitals. This pooled resource enables urgently needed real-world evidence in pediatrics, a population where trail evidence is often lacking.

Early results from the network around SPH demonstrate the registry's potential and scientific promise while highlighting key issues related to sustainability, governance, interoperability, and resourcing decisions that must guide its next phase.

References: 1.doi:10.1136/bmjopen-2024-091884

[28-P1-06] A Globally Interoperable Platform for Activity Research in Cell Therapy and Hematopoietic Stem Cell Transplantation: A WBMT Initiative

Presenter: Daniel Neumann

Authors: Daniel Neumann [1]; Helen Baldomero [2]; Dietger Niederwieser [3]

Affiliations: [1] Institute of Medical Informatics, Statistics and Epidemiology, Leipzig; [2] The Worldwide Network for Blood and Marrow Transplantation (WBMT) Activity Survey Office, University Hospital Basel, Basel, Switzerland, Basel; [3] University of Pennsylvania, Pennsylvania

Introduction

Global real-world data infrastructures in highly specialized domains remain limited by heterogeneous data models, fragmented governance, and lack of interoperability across registries and care systems. While common data models such as OMOP enable standardized analytics within defined networks, they do not address the requirements for globally coordinated, domain-specific data integration. To overcome these limitations, a federated, interoperable platform for transplant activity and outcome research has been established under the Worldwide Network for Blood and Marrow Transplantation.

Methods

The platform is based on a modular, FHIR-compliant architecture enabling semantic interoperability across heterogeneous source systems, including local documentation systems and national registries. Data are captured or transformed into standardized FHIR resources using controlled terminologies and mapping specifications. A federated execution model is implemented, in which analysis definitions are distributed to participating nodes, executed locally, and return aggregated, non-identifiable results. Governance is structured as a multi-tier model comprising global coordination, national/regional oversight, and local data stewardship. Data access is regulated through granular, purpose-bound data use agreements. Operational components include metadata-driven data harmonization, versioned analysis definitions, audit trails, and continuous data quality assessment using rule-based validation and cross-variable consistency checks.

Results

The infrastructure integrates approximately 2 million transplant and cell therapy procedures collected over more than 20 years and harmonized into a FHIR-based representation. It supports distributed, reproducible analyses across heterogeneous systems without central data pooling. In selected environments, the platform enables linkage to external datasets (e.g., incidence, health economic indicators), allowing multi-level analyses across epidemiological and system-level dimensions. The system is actively used for routine reporting and study execution.

Discussion

This platform demonstrates that globally scalable, federated data infrastructures can be operationalized through alignment of interoperable data models, distributed analytics, and enforceable governance. It provides a foundation for reproducible, privacy-preserving real-world evidence generation across jurisdictions. Future work will focus on formalizing analytical pipelines, extending outcome representations, and integrating advanced methods (e.g., AI-assisted validation and federated learning) to further enhance scalability and analytical depth.

[28-P1-07] Reproducible research infrastructure for multicenter biomedical research projects using configuration management

Presenter: Sebastian Kaletta

Authors: Sebastian Kaletta [1]; Friedemann G. Ringwald [1]; Petra Knaup [1]; Matthias Ganzinger [1]

Affiliations: [1] Institute of Medical Informatics, Heidelberg University Hospital, Heidelberg

Introduction

Reproducibility and data sovereignty are key challenges in collaborative biomedical research. Multisite consortia such as SMART-CARE, a systems medicine project investigating proteomic and metabolomic biomarkers for predicting cancer recurrence via mass spectrometry [1], require research data management applications that can be used uniformly across institutions, maintained in a traceable manner, and initialized with study-specific metadata. Ad-hoc server configurations are difficult to reproduce, test, or transfer, which impairs both operational efficiency and compliance with the FAIR principles. Using SMART-CARE as an example, we present a software-defined infrastructure approach that combines configuration management, containerization, and automated metadata loading to address these challenges.

Methods

The infrastructure is defined entirely as code, using Git for version control, Ansible for configuration management, and GitLab CI/CD pipelines for deployment, enabling a fully automated and reproducible server environment. The metadata is loaded from a CDISC Operational Data Model (ODM) file, converted into the openBIS metadata format and loaded at initialization.

Results

The defined infrastructure can be deployed on new servers with a single Ansible run, resulting in an identical, configured research environment. This includes openBIS for data management, Nextcloud for uploading large files, GitLab for the analysis teams' repositories, and R Shiny Apps in conjunction with Shinyproxy for interactive data analysis. On a second server (physically separated), Mainzelliste handles the pseudonymization of patient data. During deployment, study-specific ODM-compliant metadata is converted to the openBIS Excel format with a script and preconfigured, enabling immediate data entry without manual setup.

Discussion

This "infrastructure-as-code" approach allows the application of the FAIR principles to the research infrastructure itself: discoverability through standardized metadata structures, accessibility through dedicated web services, interoperability through ODM-compliance, and reusability through version-controlled deployment definitions. Preloading ODM-derived metadata into openBIS at startup ensures that new servers or projects operate using the same controlled vocabularies and structures. The separation of infrastructure code and operational data promotes traceability and simplifies updates. The approach is extensible to other multi-site biomedical consortia requiring uniform, reproducible research environments.

[1] F. G. Ringwald et al., "Architecture of the mass spectrometry data management pipeline in the SMART-CARE project", Stud. Health Technol. Inform., vol. 310, pp. 1016-1020, 2024.

[28-P1-08] Centralized and Structured Metadata Management as a Foundation for Interoperable Processes in Clinical Research

Presenter: Nanae Sai

Authors: Nanae Sai [1]; Andrew Heidel [1]; Henner M. Kruse [1]; Florian Rißner [1]; Kutaiba Saleh [1]; Danny Ammon [1]; André Scherag [2]

Affiliations: [1] Universitätsklinikum Jena, Jena; [2] Institut für Medizinische Statistik, Informatik und Datenwissenschaften, Universitätsklinikum Jena, Jena, Jena

Introduction

In clinical research, in addition to primary data, various types of metadata are required to ensure legally and technically compliant data. These metadata are often distributed across multiple subsystems and embedded in system-specific data models.

In the Model Project Genome Sequencing, multiple organizational units at the University Hospital Jena contribute data using heterogeneous documentation processes, data structures, and export formats [1].

To process these distinct types of data from diverse departments we developed a data management system within the Model Project Genome Sequencing to support improved diagnostics and individualized therapy decisions.

Methods

To enable centralized metadata management, a data management system using REDCap [2] was implemented. Structured input forms were developed to capture patient and sample identifiers, administrative information, and delivery-related metadata across systems. Data entry was supported by validation mechanisms that ensured completeness and indicated whether datasets are ready for transfer.

An API-based interface with token-based authentication allowed automated and secure integration with external systems, supporting data extraction, transfer, and other administrative information.

Results

This approach enables the cross-departmental integration of distributed data sources into a consistent structure, thereby reducing manual effort and minimizing potential errors in data integration and transfer.

It further supports interoperable data access through a unified API, enabling flexible use of heterogeneous data sources. In productive use, the approach demonstrated stable and scalable processing of complex clinical and genomic metadata.

Discussion

The combination of REDCap-based data capture and API-driven integration provides a flexible and scalable approach for managing heterogeneous metadata across multiple departments. Remaining challenges include mapping complexity due to differing data formats and the need for continuous adaptation of the transformation logic.

The presented approach tackles a general challenge in research data management, namely the collection and standardization of distributed and heterogeneous metadata under real-world constraints. The developed architecture can serve as a transferable model for similar data-intensive and multi-department research environments and contribute to more interoperable data infrastructures.

[1] Till A et al. Germany's national genomDE strategy. Nat Med. 2025 Dec31(12):3981-3984. doi: 10.1038/s41591-025-03991-2. PMID: 41094033.

[2] Vanderbilt University. REDCap. Available from: https://projectredcap.org

[28-P1-09] Modern health data catalogs: Evaluating data sharing capabilities by establishing the state-of-the-art

Presenter: Marcel Dilje

Authors: Marcel Dilje [1]; Kathrin Sobe [1]; Alexander Engelbrecht [1]; Mathias Rühle [2]; Matthias Reusche [2]; Eveline Prochaska [1]

Affiliations: [1] Institute for Medical Informatics and Biometry, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Dresden; [2] Leipzig Research Centre for Civilization Diseases, Faculty of Medicine, Leipzig University, Leipzig

Introduction

The application of AI and machine learning in health research requires increasing amounts of suitable medical data, often siloed within organizations and inaccessible for researchers. Data catalogs are an approach to enable cross-organizational discovery of data. Despite their widespread usage, no overview or standards for modern data catalogs exist.

Therefore, this paper aims to establish a state-of-the-art and requirements for data catalogs through a scoping review encompassing white and grey literature.

Methods

A scoping review across four scientific literature databases (N=906, n=20) was conducted, supplemented by an analysis of existing implementations consisting of data catalog products (N=110, n=20) and 40 health data catalogs.

Results

Through the review of literature and existing implementations, 40 requirements for data catalogs were extracted. A strong convergence encompassing three core requirement categories was observed, consisting of access control, data security and data discovery. The analysis of data catalog products reveal that they largely fulfill the requirements belonging to access control and data security.

Among the 40 health data catalogs, no sufficient documentation for analysis was available. The consistent lack of technical documentation outlines a structural problem in health data catalogs, preventing the examination of their core requirement implementation and creating a glaring research gap.

Discussion

The results highlight a strong discrepancy between literature and practice. Data catalog products feature a robust technical foundation yet lack the emphasized key data sharing functionalities. For health data catalogs, due to the aforementioned research gap, it makes it impossible to determine whether the core requirements, most importantly FAIR compliance and data discovery, were unmet or simply undocumented. Without sufficient documentation, further analysis and comparative evaluations are severely limited.

Based on these findings, current implementations appear insufficient to fully support data sharing functionalities. Until health data catalogs implement improved documentation and the core requirements, the efficiency of public health data exchange is severely limited, negatively impacting and slowing the application of machine learning and AI in health research. Therefore, the core requirements and research gap identified in this study are proposed as foundational reference for the development of future data sharing infrastructures. (Part of KIMed, SAB and EU founded no. 100733200)

[28-P1-10] Access and Use Processes of Health Data Platforms in Germany: A Comparative Analysis of FDZ Gesundheit and FDPG

Presenter: Johanna Apfel-Starke

Authors: Matthias Katzensteiner [1]; Johanna Apfel-Starke [2]

Affiliations: [1] Hochschule Hannover, Hannover; [2] Medizinische Hochschule Hannover Hochschule Hannover, Hannover

Introduction

The increasing availability of health data has created new opportunities for medical research while raising challenges regarding access, governance, and data protection. In Germany, two central infrastructures support research access to health data: the Research Data Centre for Health (FDZ Gesundheit) and the German Research Data Portal for Health (FDPG). While both aim to facilitate the secondary use of health data, they differ substantially in their organizational structures, access procedures, and data characteristics.

Methods

This study is based on a structured literature review complemented by the analysis of legal frameworks, institutional documents, and practice-oriented sources. A criteria catalogue was developed to enable a systematic comparison of both platforms, focusing on access procedures, governance structures, and data characteristics.

Results

The analysis reveals fundamental differences between FDZ Gesundheit and FDPG. FDZ Gesundheit follows a centralized model, providing access to pseudonymized claims data within a secure processing environment (SPE), ensuring high data protection standards. In contrast, FDPG is based on a federated infrastructure, where data remain at local university hospitals and are accessed via Data Integration Centers (DICs), allowing detailed clinical analyses. These structural differences result in varying levels of data coverage, granularity, and access complexity.

Discussion

Both platforms address different research needs and should be understood as complementary rather than competing infrastructures. FDZ Gesundheit enables large-scale population-based analyses, while FDPG supports detailed clinical research. However, challenges remain regarding interoperability, legal harmonization, and efficient access procedures. The findings highlight the importance of coordinated development of health data infrastructures to fully realize the potential of data-driven research.

[28-P1-11] Overcoming OMOP Compatibility Barriers in Clinical Research Data Migration for Real-World Evidence Studies: Integrating the OMOP CDM via Data2Evidence

Presenter: Johanna Apfel-Starke

Authors: Norman Schönfeld [1]; Nicole Hechtel [2]; Peter Hoffmann [3]; Tim Walz [4, 3]; Johanna Apfel-Starke [1]; Nicole Bönisch [1]

Affiliations: [1] Hannover Medical School, MHH Information Technology Department, Hannover; [2] Hannover Medical School, Peter L. Reicherts Institute for Medical Informatics, Hannover; [3] Data4Life Asia Limited, Singapore; [4] D4L data4life gGmbH, Potsdam

Introduction

Standardized data models, such as the Observational Health Data Sciences and Informatics (OHDSI) Common Data Model (OMOP CDM), are critical for enabling multi-site real-world evidence (RWE) studies. However, migrating clinical data into OMOP CDM often encounters infrastructural and compatibility challenges, impeding seamless adoption and analysis.

Background: The Hannover Medical School (MHH) Enterprise Clinical Research Data Warehouse (ECRDW) [1], established in 2013 and housing data from over 2.6 million patients (including 20 million diagnoses and 3.4 billion data points as of March 2026. [1] Within the EHDEN project (2022-2023) [2], MHH initiated the transformation of its ECRDW data into OMOP CDM. A proof-of-concept using 100,000 patient records was implemented on PostgreSQL, but revealed structural mapping complexities, data-type mismatches, and performance limitations. Migration to Microsoft SQL Server (MSSQL) on Linux significantly improved ETL efficiency through SSIS automation and resolved performance bottlenecks. However, these incompatibilities with standard OHDSI analysis tools (e.g., ATLAS, Achilles, R study scripts), conflicting with OHDSI's philosophy of a unified analytical environment. Data quality validation using Achilles and the DataQualityDashboard confirmed the readiness of the dataset, highlighting the need for solutions that preserve analytical compatibility while supporting flexible infrastructures.

Methods

To address this gap, Data4Life developed the open-source Data2Evidence platform [3], providing a user-friendly interface for data integration, cohort definition, and analysis. In this project, the MSSQL-based OMOP CDM instance at MHH is integrated into Data2Evidence, enabling low-barrier cohort queries and interactive analyses without requiring modifications to OHDSI analysis scripts.

Results

The platform uses modular adapters and an internal query engine to bridge database-specific implementations with OMOP-compliant workflows. Additionally, it is planned zu evaluate its capability for multi-CDM integration, including the transformation of openEHR data into OMOP CDM. We therefore show preliminary results.

Discussion

By harmonizing the flexibility of the infrastructure with OHDSI-compliant analytics, Data2Evidence aims to eliminate the trade-off between performance and interoperability. Future work will evaluate Data2Evidence's efficacy in reducing analysis setup time, enhancing cross-CDM integration scalability, and supporting broader adoption of standardized RWE frameworks.

References

[1] https://doi.org/10.1007/978-3-030-06016-9_17

[2] https://doi.org/10.3233/SHTI230485

[3] https://github.com/OHDSI/Data2Evidence

[28-P1-12] Development and Application of a Structured Piloting Concept for the Implementation and Migration of Modules of the MII Core Data Set in Data Integration Centers

Presenter: Danny Ammon

Authors: Danny Ammon [1]; Karoline Buckow [2]; Jördis Beck [3]; Romina Blasini [3]; Regina Moritz [4]; Holger Stenzhorn [4]

Affiliations: [1] Universitätsklinikum Jena, Jena; [2] TMF - Technology and Methods Platform for Networked Medical Research e.V., Berlin; [3] Justus-Liebig-University Gießen, Gießen; [4] Data Integration Centre, Saarland University Medical Center, Homburg/Saar, Homburg

Introduction

The Medical Informatics Initiative's Core Data Set (MII-CDS) [1] is a national standard for sharing clinical treatment data across healthcare sites. Organized into modules with interoperable data structures managed by Data Integration Centers (DIC), the evolving CDS specifications regularly introduce new or updated modules. This creates recurring challenges for DIC in implementation, testing, and migration. We present a structured piloting concept designed to facilitate a coordinated and efficient rollout of these CDS modules.

Methods

The concept is being developed as part of the research infrastructure funding of the Data Integration Centers in the Network of University Medicine (NUM-DIZ) [2]. In several structured cross-site coordination rounds, web conferences and written revisions, the project partners jointly developed a process model for the introduction and updating of CDS modules. Typical work processes of the DIC were considered, like the development and adaptation of ETL processes (extract - transform - load) for the integration of local clinical data sources.

Results

The pilot concept developed distinguishes between three central processes: new implementation, piloting as first implementer, and migration to an updated module. Structured steps are described for each procedure, including familiarization with specifications, analysis of local data sources, mapping of data structures, ETL implementation, and testing. The practical testing of the concept initially took place in 2026 on three pilot DIC. The microbiology module (intermediate version 2026) and an updated version of the laboratory module are being piloted.

Discussion

The presented concept offers a structured framework for introducing new or updated CDS modules and addresses key challenges of cross-site implementation. Systematic piloting at selected DIC allows implementation experience to feed back into further specification development early on. Over time, this process can help reduce implementation efforts, make migrations more predictable, and support consistent MII CDS implementation in the DIC. Piloting within the NUM-DIZ infrastructure also strengthens the integration of specification development with practical implementation.

References:

[1] Medical Informatics Initiative Germany (ed.) The Medical Informatics Initiative's Core Data Set. 2026. Available from: https://www.medizininformatik-initiative.de/en/medical-informatics-initiatives-core-data-set [cited 31 Mar 2026]

[2] Network of University Medicine (ed.) NUM Data Integration Centers (NUM-DIZ). 2026. Available from: https://www.netzwerk-universitaetsmedizin.de/en/platforms/num-diz [cited 31 Mar 2026]

[28-P1-13] Semantic Annotation and Data Transformation of Internal Microbiology Codes into International Code Systems within the MII Microbiology FHIR Profiles

Presenter: Andrew Heidel

Authors: Andrew Heidel [1]; Stefan Glöckner [1]; Danny Ammon [2]

Affiliations: [1] University Hospital Jena, Jena; [2] Universitätsklinikum Jena, Jena

Introduction

Producing interoperable FHIR resources from healthcare data remains, more than 10 years after the first FHIR standards, a formidable task due to various factors including semantic annotation. However, it is critical to have microbiology data in an interoperable format since it is used not only for patient care and research, but also for public health monitoring and potentially to direct government policy [1].

Our goal was to generate microbiology FHIR resources [2] from internal laboratory information system (LIS) HL7 messages. This transformation requires accurate semantic annotation and translation into the codes needed. To increase accuracy, our annotation process was performed with coordination between the Institute for Medical Microbiology and the Data Integration Center within the Jena University Hospital.

Methods

It was determined that the LIS data could be divided into four main areas, culture results with bacterial count information, antimicrobial susceptibility testing, microscopy results, and molecular/immunological tests. These data could then be used to produce 10 different types of microbiological FHIR profiles. The profiles required the following codes to be annotated and translated: a main code that describes the test or what will be identified, the result code or the units of the result if the result

is quantitative, a sometimes-optional method code and a sometimes-required component element with extra information.

Results

We were able to annotate all codes for six profiles, and mostly annotate codes for four additional profiles where most codes could be mapped, but a small number of codes presented problems. The problematic cases either had no LOINC code that accurately matched the meaning of the internal code or had a result format that did not match the current FHIR profile result format.

Discussion

We were able to annotate most codes needed to create microbiology FHIR resources and thereby reliably captured the medical meaning of the original source data.

References:

1. Rutte EA, et al. A case for ongoing structural support to maximise infectious disease modelling efficiency for future public health emergencies: A modelling perspective. Epidemics. 202446:100734.

2. Rinaldi, E, et al. Towards interoperability in infection control: a standard data model for microbiology. Sci Data. 202310:654.

[28-P1-14] Navigating Technical Challenges in FHIR-Based Patient Reported Outcome Measures: Towards Interoperability with the MII PRO Core Dataset

Presenter: Yannik Warnecke

Authors: Yannik Warnecke [1]; Dominik Heider [2]; Michael Storck [3]

Affiliations: [1] University of Münster, Münster; [2] Institute of Medical Informatics, University of Münster, Münster; [3] Universität Münster, Münster

Introduction

Patient-Reported Outcome Measures (PROMs) are vital for capturing subjective clinical data, yet electronic PROM tools often suffer from fragmentation and siloed data storage. To facilitate cross-site data sharing, large-scale interoperability must be integrated into these digital tools.

Methods

We developed a version-agnostic HL7 FHIR system for Mobile Patient Survey (MoPat), designed to integrate PROMs into various FHIR-based infrastructures. To evaluate the system, we used the Medical Informatics Initiative (MII) PRO core dataset as a use-case. This evaluation aimed to assess usability for MII projects and identify technical or functional gaps.

Results

While technical integration was successful, semantic annotation remains a significant hurdle. Specifically, importing PRO questionnaires from the MII core dataset resulted in data loss concerning score calculation logic and specific code systems.

Discussion

The results demonstrate the potential of utilizing standardized tools for both data collection and structuring, as they significantly reduce development time while enhancing data standardization. However, it simultaneously reveals tension between highly regulated data collection software and research-focused data structures that prioritize more granular detail. This study validates the potential of electronic PROM tools to adopt standards like the MII core dataset. Future development should prioritize the inclusion of standardized code systems to further enhance semantic interoperability.

[28-P1-15] A User-Centered, SMART-on-FHIR-Based Dashboard Prototype for Diabetic Macular Edema

Presenter: Katja Hoffmann, Hamed Mishian

Authors: Katja Hoffmann [1]; Sophia Grummt [2]; Hamed Mishian [2]; Maximilian Nussbaumer [3]; Markus Wolfien [2, 4]; Gabriel Stolze [5]; Katrin Engelemann [5]; Martin Sedlmayr [6]

Affiliations: [1] Zentrum für Medizinische Informatik / Institut für Medizinische Informatik und Biometrie (IMB) Medizinische Fakultät Carl Gustav Carus, Dresden; [2] Institute for Medical Informatics and Biometry, Faculty of Medicine Carl Gustav Carus, TUD Dresden University of Technology, Dresden; [3] Max Health Inc., San Francisco; [4] Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI), Dresden; [5] Department of Ophthalmology, Klinikum Chemnitz gGmbH, Chemnitz; [6] TU Dresden, Dresden

Introduction

Diabetic macular edema (DME) management requires the integration of ophthalmic and systemic clinical data (HbA1c, hs-CRP, blood pressure), yet relevant information often remains fragmented across separate, poorly interoperable systems. A further barrier is that widely deployed FHIR servers, such as Blaze, which is broadly used in the German Medical Informatics Initiative (MII) infrastructure, do not natively provide SMART-on-FHIR support (e.g., SMART App Launch, OAuth2 authorization endpoints, PKCE, launch context, and scopes), making standardized, secure app access difficult in practice. Building on a previously published User-Centered Design (UCD) process, this study focuses on designing and implementing an interoperable dashboard prototype that consolidates ophthalmic and systemic parameters in a single, vendor-neutral view while addressing this authorization gap.

Methods

The prototype was developed using an agile, minimal-viable-product approach within a layered, SMART-on-FHIR-compatible architecture comprising data provisioning (Blaze), identity and access management (Keycloak), a Proxy SMART middleware adding SMART-on-FHIR authorization to a server lacking native support, and a modular web-based dashboard. Ophthalmic and systemic clinical data were standardized using HL7 FHIR by adopting the MII core data set and the EyeMatics core data set, while project-specific FHIR profiles were defined where necessary. As the UCD process has been reported previously, the present work focuses on the technical realization of the architecture and data integration.

Results

The prototype demonstrates how ophthalmic and systemic clinical data can be integrated using standardized FHIR resources and a decoupled authorization architecture that retrofits SMART-on-FHIR onto a natively unsupported FHIR server. Color-coded thresholds and trend indicators provide basic visual guidance, and missing values are made explicit. The implementation, deployment configuration, and synthetic test data are released as open source.

Discussion

The dashboard prototype demonstrates how standards-based interoperability and a user-centered interface design can consolidate heterogeneous ophthalmic and systemic clinical data in a vendor-neutral manner, including on servers without native SMART-on-FHIR support. It provides a technical foundation for interoperable clinical applications by enabling standardized, secure access to integrated clinical data and their visualization. Future work will focus on extending the prototype with advanced decision support models, evaluation in clinical workflows and regulatory assessment under the Medical Devices Regulation.

[28-P1-16] International Patient Summary in Street Medicine

Presenter: Philipp Plicka

Authors: Philipp Plicka [1]; Alfred Winter [1]; Franziska Jahn [2]

Affiliations: [1] Universität Leipzig, Leipzig; [2] Institut für Medizinische Informatik, Statistik und Epidemiologie, Universität Leipzig, Leipzig

Introduction

People experiencing homelessness face substantial barriers to primary healthcare, resulting in fragmented care, high disease burden and increased mortality. Street medicine offers low-threshold, outreach-based care in patients' environments [1], but is limited by insufficient access to patient information and a lack of standardized documentation. The absence of structured, interoperable data restricts continuity of care and integration with regular healthcare. Interoperability standards such as the International Patient Summary (IPS) in HL7 FHIR [2] offer potential solutions by enabling portable, structured health information across care settings.

Methods

This study follows a multi-stage approach combining requirements analysis, specification, and prototypical implementation. Preliminary studies, documentation practices, and empirical insights from workshops and semi-structured stakeholder interviews were systematically consolidated to identify requirements for a street medicine documentation system. These were formalized, structured, and prioritized in a comprehensive specification. A market analysis of open-source electronic health record systems was conducted and evaluated using a criteria-based scoring model, resulting in the identification of a suitable software platform. Based on this, a functional prototype was implemented using open-source technologies. It supports longitudinal electronic patient records and standarized

data collection. Data elements were accessible as FHIR resources via REST interfaces. An integration layer was developed to aggregate and transform these data into an IPS document.

Results

The work delivers (1) a structured and validated requirements specification for electronic documentation systems in street medicine, (2) a transparent evaluation and selection of an open-source EHR, and (3) a functional prototype demonstrating the feasibility of structured, longitudinal documentation and interoperability. Patient data were successfully extracted as FHIR resources, aggregated, and transformed into a valid IPS document, showing that interoperable and standards-compliant data exchange is technically achievable.

Discussion

The results demonstrate that combining requirements-driven system design with open standards enables a practical pathway toward interoperable documentation in street medicine. By linking specification, system selection, and prototypical implementation, the work moves beyond conceptual proposals and provides a reproducible technical foundation. Remaining challenges include real-world validation, usability, and workflow integration, which should be addressed in future research and implementation efforts.

[1] Street Medicine Institute (n.D.). About SMI. https://www.streetmedicine.org/about-smi.

[2] HL7 International (2025). International Patient Summary ImplementationGuide. https://hl7.org/fhir/uv/ips/.

[28-P1-17] Promoting cross-project interoperability in telemedicine innovation clusters: A dissemination and implementation approach

Presenter: Christian Otto, Cord Spreckelsen

Authors: Christian Otto [1]; Cord Spreckelsen [1]

Affiliations: [1] Universitätsklinikum Jena, Jena

Introduction

The wecare alliance brings together over 20 projects under a cluster-based funding scheme with the aim of improving regional healthcare through modern, often sensor-based telemedicine applications. To avoid IT fragmentation among the individual projects, measures to establish common interfaces and promote cross-project interoperability were initiated at an early stage in the wecare-AM starter project.

Methods

Central to this is a hub-and-spoke metaphor, which prioritises a needs-based, interface-driven data exchange over centralised data storage. Established FHIR-based [1] standards such as the MII Core

Set, ISiK and Medical Information Objects (MIOs) serve as the foundation. Methodologically, wecare-AM combines an agile approach with consultancy and training services, as well as a technical platform to promote FHIR skills. The core component of this platform is a FHIR Readiness Module that teaches the development and testing of key FHIR concepts.

Results

wecare-AM carried out needs analyses and offered cross-project consultancy, workshops and training sessions. In addition, the FHIR Readiness module was implemented, based on a containerised microservice architecture with a HAPI FHIR [2] server as the central component. Supporting services ensure authentication, data storage, session management and secure communication. To facilitate skills development, practical, automatically assessable Jupyter Notebook-based learning materials were developed. Furthermore, role and user management as well as a logging system were implemented, enabling a transparent assessment of progress ('FHIR readiness') within the alliance.

Discussion

The combined approach of technical infrastructure, consultancy and training has proven to be flexible and resource-efficient in a heterogeneous project environment. Particular emphasis should be placed on promoting the dissemination of interoperable solutions through cross-project training and support services. The introduced scoring system provides additional incentives for implementation. A key challenge, however, is that many projects did not initially allocate sufficient resources for interoperability, meaning that its implementation depends largely on the commitment of those involved.

Funded as part of the "WIR! - Wandel durch Innovation in der Region" funding scheme by the German Federal Ministry of Research, Technology and Space (funding reference: 03WIR5804)

[1] Index - FHIR v5.0.0 [Internet]. [cited 2026 Mar 31]. Available from: https://www.hl7.org/fhir/

[2] hapifhir/hapi-fhir [Java] [Internet]. HAPI FHIR 2026 [cited 2026 Mar 31]. Available from: https://github.com/ha
fhir

[28-P1-18] Hierarchical and LLM-assisted identification of semantically similar data elements in heterogeneous diabetes datasets using OMOP CDM vocabulary

Presenter: Ivona Anastasova

Authors: Ivona Anastasova [1]; Kirstin Tümler [1]; Angela Dedié [1]; Lars Oest [1]; Andreas Hungele [1, 2]; Alexander Eckert [1, 2]; Stefanie Lanzinger [1, 2]

Affiliations: [1] German Center for Diabetes Research e.V. (DZD), Neuherberg; [2] Institute of Epidemiology and Medical Biometry, University of Ulm, Ulm

Introduction

The movement toward large-scale secondary research, driven by initiatives like the Observational Health Data Sciences and Informatics (OHDSI) network and the European Health Data Space, necessitates robust methods to identify semantically similar data elements across heterogeneous health data. To address this need, we present a hierarchy-based matching approach using the OHDSI

ATHENA vocabulary, enhanced with large language models. We aimed to extract common items between two diabetes-focused datasets which had already been independently mapped to the OMOP common data model (CDM): (1) the core dataset of the German Center for Diabetes Research (DZD), and (2) items from the diabetes prospective follow-up registry (DPV).

Methods

Non-standardized OMOP concepts were mapped to their standard representation to maximize identification of concordant concepts. Substances annotated with RxNorm were mapped to Snomed CT. Hierarchical concept matching was performed in RStudio. The ATHENA CONCEPT_ANCESTOR and CONCEPT_RELATIONSHIP table were used to identify equivalent and parent-child concept pairs. ChatGPT was used to screen remaining unmatched concepts, and identified pairs were manually validated. Subsequently unmatched concepts were manually assessed. We report the number of identified pairs, and the number of unique concordant concepts per dataset to account for one-to-many hierarchical mappings.

Results

After excluding non-clinical and questionnaire items, 222 DPV and 177 DZD OMOP concepts were evaluated. In total, 152 concordant pairs were identified, comprising 105 unique DPV concepts (47%) and 70 unique DZD concepts (40%) of the respective datasets.

Hierarchical matching identified 120 pairs, including 15 equivalent and 105 parent-child pairs. Screening with ChatGPT identified 22 correct and 2 incorrect pairs, while preliminary manual review of the remaining concepts yielded 10 pairs. All 152 pairs span several OMOP domains: condition (n=99), measurement (n=34), observation (n=11), drug (n=2), procedure (n=1), mixed domains (n=5).

Discussion

By leveraging systematic relationships within the ATHENA vocabularies, our approach enhances precision and transparency compared with purely LLM-driven methods and enables hierarchical matching beyond the scope of the source vocabularies. Key challenges include heterogeneous mapping strategies, incomplete hierarchies, and semantic abstraction introduced by the non-standard-to-standard-mapping. This approach provides a reproducible and scalable framework for identifying common items and research questions across datasets.

[28-P1-19] From Silos to Ecosystems: A Systemic Innovation Framework for Integrated Clinical Data Management in ADPKD Care

Presenter: Oliver Diekmeier, Lada Liubisheva

Authors: Oliver Diekmeier [1]; Lada Liubisheva [2]; Oliver Stremel [2]; Jan Schroeder [2]; Alina Troglio [2]; Malte Bartram [2]; Nina Rank [3]; Jan Halbritter [3, 4]; Müller Roman-Ulrich [5]; Oya Beyan [1]

Affiliations: [1] Universität zu Köln, Köln; [2] University Clinic of Cologne, Köln; [3] Charité - Universitätsmedizin Berlin, Berlin; [4] niversity Medical Center Freiburg, Freiburg; [5] Universitätsklinikum Düsseldorf, Düsseldorf

Introduction

The DICE-CD project, which launched in January 2026, addresses the persistent challenge of fragmented healthcare data. Clinical information is frequently trapped in isolated systems, preventing an end-to-end (E2E) data flow across the patient journey - from diagnosis to long-term follow-up. Using Autosomal Dominant Polycystic Kidney Disease (ADPKD) as a complex multicenter use case, the project aims to transform disparate data silos into a networked ecosystem. The primary goal is to enhance diagnostic precision and operational efficiency by bridging the gap between technical infrastructure and clinical work practice.

Methods

The methodology follows a Design Science Research approach, integrating technical development with organizational transformation. The initial phase focuses on two core work packages: Clinical Process and Patient Path Modeling (WP1): Multi-professional teams at three university hospitals (Cologne, Düsseldorf, Berlin) utilize qualitative interviews and workshops to map current workflows into formal BPMN-based models. This identifies critical data touchpoints and cognitive load factors to derive user-centric requirements. Data Ecosystem Analysis (WP2): A systematic inventory of the existing IT landscape (e.g., HIS, PACS, LIS) is conducted. Systems are evaluated regarding syntactic interoperability (utilizing the HL7 FHIR standard) and semantic interoperability (utilizing SNOMED CT and LOINC), alongside data quality and security assessments.

Results

As a newly established initiative, the project targets the development of a validated Systemic Innovation Framework, consisting of: A Reference Architecture (Blueprint) for clinical data ecosystems to support the seamless integration of laboratory, imaging, and genetic data. The Clinical Data Usage Maturity Model, a framework to measure and steer digital maturity within clinical processes. User-centric prototypes, including a clinical dashboard and predictive AI-based modules for therapy decision support in ADPKD.

Discussion

DICE-CD focuses on real-time clinical integration, aiming to embed digital innovation directly into the daily routine of medical staff. By systematically linking technical architecture with human-machine interaction and organizational change management, the project provides a scalable methodology for digital transformation. This holistic approach is designed to be transferable to various medical domains beyond nephrology.

[28-P1-20] Less Knowledge Without Exchange: Experiences in Developing a Data Sharing Framework Process Plugin for EyeMatics - Lessons Learned

Presenter: Sascha Prus

Authors: Sascha Prus [1]; Seyedeh Delaram Mirraziroudsari [2]; Daniel Schröder [1]; Johannes Oehm [3]; Lothar Dickhoff [4]; Thomas Rüland [5]; Florisa Zanier [5]; Sebastian Berthe [6]; Seyed Reza Mazlooman [7]; Lucian Paulet [6]; Christopher Hampf [8]; Eva Kobak [9]; Hendrik Felderhoff [10]; David Heines [11]; Myriam Lipprandt [12]; Andreas Stahl [13]; Dagmar Waltemath [6]; David Merle [14]; Oliver Kohlbacher [15, 16, 17]; Philipp Berens [18]; Marius Ueffing [19]; Focke Ziemssen [20]; Thomas Wendt [21]; Vinodh Kakkassery [22]; Frank Nüßler [23]; Alexander Schuster [24]; Katja Hoffmann [25]; Eter Nicole [26]; Tobias Brix [1]

Affiliations: [1] Institute of Medical Informatics, University of Münster, Münster; [2] Institute of Medical Informatics - Research for AI-supported, personalised medicine, University of Münster, Münster; [3] Universität Münster, Münster; [4] Uniklinik RWTH Aachen, Aachen; [5] Data Integration Center, University Hospital RWTH Aachen, Aachen; [6] Universitätsmedizin Greifswald, Greifswald; [7] Core Unit Data Integration Centre, University Medicine Greifswald, Greifswald; [8] Institut für Community Medicine, Universitätsmedizin Greifswald, Greifswald; [9] Universität Tübingen, Tübingen; [10] Medical Data Integration Center, Tübingen University Hospital, Tübingen; [11] Klinik für Augenheilkunde, Rheinisch-Westfälische Technische Hochschule Aachen, Aachen; [12] Medical Faculty, RWTH University, Aachen, Aachen; [13] Department of Ophthalmology, University Medicine Greifswald, Greifswald; [14] Department für Augenheilkunde, Universität Tübingen, Tübingen; [15] Interfaculty Institute for Biomedical Informatics (IBMI), University Tübingen, Tübingen; [16] Department of Computer Science, Applied Bioinformatics, University of Tübingen, Tübingen; [17] Institute for Translational Bioinformatics, University Hospital Tübingen, Tübingen; [18] Hertie Institute for AI in Brain Health, University of Tübingen, Tübingen; [19] Institute for Ophthalmic Research, Department for Ophthalmology, University of Tübingen, Tübingen; [20] Department of Ophthalmology, University of Leipzig, Leipzig; [21] Data Integration Center, University of Leipzig Medical Center, Leipzig; [22] Department of Ophthalmology, Clinic Chemnitz, Chemnitz; [23] Klinikum Chemnitz gGmbH, Chemnitz; [24] Department of Ophthalmology, Johannes Gutenberg University Mainz, Mainz; [25] Zentrum für Medizinische Informatik / Institut für Medizinische Informatik und Biometrie (IMB) Medizinische Fakultät Carl Gustav Carus, Dresden; [26] Department of Ophthalmology, University Münster, Münster

Introduction

The Medical Informatics Initiative (MII) has spawned a relevant multi-centric use-case in ophthalmology, namely the EyeMatics (Eye disease "treated" with interoperable medical informatics) project. An important step within EyeMatics is the development of a process plugin for the Data Sharing Framework (DSF), enabling the exchange of research data across participating institutions. State of the art: The DSF is an established component of the MII infrastructure for secure data exchange. Common methods include the pseudonymization of consented data and privacy-preserving record linkage.

Methods

Familiarization with regulatory requirements and the technical infrastructure necessary for developing a DSF process plugin. Implementation: The implementation of the DSF process plugin was initially based on existing functionalities that could be adapted with minor modifications. Processes, tasks, and data storage are clearly separated within the architecture. The plugin's functionalities were

tested locally and across multiple sites using test data. In addition, compatibility with real-world data (RWD) was also evaluated.

Results

The secure exchange of RWD must comply with legal requirements and overcome technical barriers. These challenges can be addressed by utilizing the infrastructure provided by the MII. Testing at multiple levels increased practical experiences and supported the continuous improvement of the DSF process plugin. The development of a DSF process plugin enables the exchange of ophthalmic RWD.

Discussion

Future iterations will place a stronger focus on real-time exchange.

The full paper pertaining to this abstract will be published in the GMDS journal "Medizinische Informatik, Biometrie und Epidemiologie (MIBE)."

[28-P1-21] Caring for Sharing: Development of an Out-of-the-Box Project Generator for the Data Sharing Framework

Presenter: Sascha Prus

Authors: Sascha Prus [1]; Yannik Warnecke [1]; Tobias Brix [1]; Michael Storck [1]; Dominik Heider [1]

Affiliations: [1] Institute of Medical Informatics, University of Münster, Münster

Introduction

The Data Sharing Framework (DSF) is an important part of the digital infrastructure of the Medical Informatics Initiative (MII) in Germany, which enables secure multicentric research [1]. However, the support of diverse use-cases across different sites results in a steep learning curve for the utilization and development of process plugins for the framework. As contribution to the DSF-community we introduce an application that streamlines the setup process by generating a local development environment with all necessary configurations.

Methods

The DSF tutorial process served as the basis for the application. The analysis revealed that DSF process development environments consist of three steps: 1. Instantiation of project files and folders 2. Initialization of development packages for the process plugin 3. Preparation of the development test setup. The application is written in Java and uses the Mustache framework for dynamic generation of files. A Graphical User Interface (GUI) is created with JavaFx.

Results

The application utilizes a template-driven backend which generates the development environment based on user input in a GUI. This ensures consistency and flexibility across projects of varying

scopes and provides a simple way to extend the tool's functionality in downstream tasks. It is possible to create a project-specific DSF development environment in short time. End users can simulate environments for a use-case of their process plugin. As initial starting point, the DSF "Hello World" process plugin is provisioned by the application, and changes can directly be re-deployed.

Discussion

The tool "DSF Project Generator" (https://github.com/imi-ms/dsf-project-generator) we developed allows developers to focus more on creating DSF process plugins and less on boundary conditions. However, further improvement considerations are needed, such as the export of project definitions for sharing as well as automation via command line tools. Use-case specific data exchange must be supported by the development of appropriate supporting tools. Contribution to the DSF- community encourages user-comfort, which may lead to a wide-spread adoption of the DSF in research.

References

[1] Hund H, Wettstein R, Heidt CM, Fegeler C. Executing Distributed Healthcare and Research Processes - The HiGHmed Data Sharing Framework. Stud Health Technol Inform. 2021 May 24278:126-133. doi: 10.3233/SHTI210060. PMID: 34042885.

[28-P1-22] A Simplified Process Plugin API for the Data Sharing Framework

Presenter: Hauke Hund

Authors: Hauke Hund [1]; Jan Böhringer [1]; Maximilian Kurscheidt [1]; Mathias Rühle [2]; Simon Tobias Schweizer [1]; Reto Wettstein [3]; Christian Fegeler [1]

Affiliations: [1] GECKO Institute, Heilbronn University of Applied Sciences, Heilbronn; [2] Leipzig Research Centre for Civilization Diseases, Leipzig University, Leipzig; [3] Institute for Medical Informatics, Heidelberg University Hospital, Heidelberg

Introduction

The Data Sharing Framework (DSF) enables secure, federated execution of distributed healthcare and research processes within the Medical Informatics Initiative and other research consortia. Since its initial open-source release [1], the framework has continuously evolved to address growing requirements in distributed data sharing. With the DSF 2 release, we introduced major architectural and developer-facing improvements based on feedback from different use-cases and research consortia. This abstract describes the design decisions and resulting benefits for plugin development and system maintainability.

Methods

We designed and implemented the Process Plugin API v2 as part of the DSF 2.0.0 release, available on GitHub [2]. The redesign focused on reducing coupling to the workflow engine, improving abstraction layers, and integrating commonly required functionality into reusable services. Backward compatibility and developer usability were key design goals, supported by migration mechanisms and improved tooling.

Results

The migration of the integrated workflow engine from Camunda 7 to Operaton 1 enabled long-term independence while preserving compatibility. A bytecode rewriting layer allows existing plugins to continue operating without recompilation.

The Process Plugin API v2 removes direct workflow engine dependencies and introduces a streamlined programming model. Integrated services, covering encryption/decryption, compression, certificate handling, logging, and FHIR server connection management, were designed to replace recurring custom implementations. The new connection API enables secure reuse of credentials across plugins using password, mTLS, and OIDC authentication.

In addition, we introduced standardized metadata handling, runtime access to plugin definitions, and the dsf-maven-plugin to automate build processes and documentation generation.

Discussion

With DSF 2, we aimed to improve both the sustainability of the platform and the developer experience. The Process Plugin API v2 establishes a stable and extensible foundation for future developments while maintaining compatibility with existing deployments. By lowering complexity and consolidating common functionality, the redesign facilitates faster and more reliable implementation of distributed data-sharing workflows.

Citations

[1] Hund H, Wettstein R, Heidt CM, Fegeler C. Executing Distributed Healthcare and Research Processes - The HiGHmed Data Sharing Framework. Stud Health Technol Inform. 2021278:126-133. doi:10.3233/SHTI210060

[2] Hund H, Wettstein R, Contributors of the DSF. Data Sharing Framework (DSF). Version 2.0.2. Published February 10, 2026. Available at: https://github.com/datasharingframework/dsf

[28-P1-23] A Standards-Based Infrastructure for Federated AI Inference in Medical Imaging

Presenter: Kaushik Manjunatha, Nimesh Singh

Authors: Obioma Pelka [1, 2]; Nicolas A. Kamzol [2]; Kaushik Manjunatha [3]; Nimesh Singh [3]; Jazia Omeirat [2]; Felix Nensa [4]

Affiliations: [1] Universitätsmedizin Essen, Essen; [2] Central IT Department, Data Integration Center, University Hospital Essen, Essen; [3] Institute for Artificial Intelligence in Medicine (IKIM), University Hospital Essen, Essen; [4] Institut für künstliche Intelligenz in der Medizin (IKIM), Universitätsklinikum Essen, Universität Duisburg-Essen, Essen

Introduction

Integrating artificial intelligence (AI) into clinical workflows requires secure, interoperable infrastructures that connect heterogeneous clinical data systems while respecting governance and regulatory constraints. Medical imaging infrastructures are built around PACS using the DICOM standard, but

vendors implement DICOM and DICOMweb with significant variation, creating interoperability challenges insufficiently addressed by the standards alone.

Methods

This work presents a standards-based infrastructural approach for federated AI inference in medical imaging, centered on a DICOMweb adapter that mediates between heterogeneous clinical PACS implementations and federated AI services. The adapter addresses real-world PACS heterogeneity through protocol translation between DICOMweb and DIMSE and configurable normalization of vendor-specific deviations, with explicit extension points for governance services such as de-identification. The approach was evaluated through clinical deployment at a university hospital and a reproducible multi-PACS simulation involving two independent Orthanc backends, together with a comparative benchmark against Orthanc's DICOMweb plugin.

Results

The comparative benchmark revealed that existing tools cannot proxy DICOMweb to remote DIMSE-based PACS and fail to support the WADO-RS multipart content type required by AI pipelines. The adapter showed consistent performance across heterogeneous backends (median QIDO-RS latency 46-50 ms, 100% success rate), and the two-week clinical deployment handled realistic load (6,061 requests, 100% success rate).

Discussion

The results confirm the necessity of a dedicated mediation layer for federated AI workflows in clinical imaging, showing that standards-based protocol translation and configuration-driven normalization provide vendor-neutral imaging access that existing open-source tools do not.

The full paper pertaining to this abstract will be published in the GMDS journal "Medizinische Informatik, Biometrie und Epidemiologie (MIBE)."

[28-P1-24] Versioning of AI Algorithms in the Open Medical Inference (OMI) Implementation Guide

Presenter: Stefan Sigle

Authors: Stefan Sigle [1]; Georg Mathes [1]; Patrick Werner [1]

Affiliations: [1] MOLIT Institute gGmbH, Heilbronn

Introduction

Radiology is a driver for innovation of artificial intelligence (AI) applications in medicine. Deployments are often localized, confined to proprietary silos, or commercially restricted, hindering multicentric adoption in clinical and research settings. The Open Medical Inference (OMI) project seeks to democratize access to AI algorithms beyond organizational borders [1]. By adhering to FAIR principles, OMI leverages existing healthcare data standards like FHIR and DICOM to facilitate integration and transparency in heterogeneous environments.

Methods

Through screening of algorithms from project partners, we established a comprehensive, structured representation of inference algorithms. This involved mapping and profiling conceptual specifications, including input and output parameters, to the FHIR Standard [2]. This work considers modelling operational aspects like version control for algorithms, services and inference endpoints, while accounting for organizational responsibility for algorithm stewardship and instance management, covering critical governance requirements.

Results

The OMI specification enables efficient versioning of algorithms and services. The utilization of common terminologies ensures semantic interoperability, allowing disparate systems to exchange data with precision and reducing manual effort required for integration.

Discussion

Integrating FHIR concepts with DICOM value sets enables a FAIR representation of AI algorithms. While OMI focuses on radiology, input and output parameters for images and metadata are defined in an abstract way. Thus, this Implementation Guide could serve as blueprint for algorithms of other clinical domains in the future. By decoupling algorithm logic from hardware constraints, OMI promotes reproducibility and facilitates validation studies across diverse populations, potentially accelerating the translation of AI from research prototypes to clinical practice and enhancing generalizability. As a result patient outcomes could be improved and the barrier to entry for clinical research could be reduced overall.

References

[1] Pelka O, Sigle S, Werner P, Schweizer ST, Iancu A, Scherer L, et al. Democratizing AI in Healthcare with Open Medical Inference (OMI): Protocols, Data Exchange, and AI Integration. RöFo - Fortschritte Auf Dem Geb. Röntgenstrahlen Bildgeb. Verfahr. 198 (2026) 173-184. doi:10.1055/a-2651-6653.

[2] Sigle S, Werner P, Schweizer S, Caldeira L, Hosch R, Dyrba M, et al. Bridging the Gap Between (AI-) Services and Their Application in Research and Clinical Settings Through Interoperability: the OMI-Protocol. (2024). doi:10.34657/13458.

[28-P1-25] Recurring Challenges in Laboratory Material Management in Biomedical Research: Lessons from menoci Development

Presenter: Sabine Andrea Smolorz

Authors: Sabine Andrea Smolorz [1]; Katja Saldeitis [1]; Robin Max Giese [1]; Christian Henke [1]; Sara Yasemin Nussbeck [1, 2]; Ulrich Sax [1, 3]; Harald Kusch [1, 3, 4]

Affiliations: [1] University Medical Center Göttingen, Department of Medical Informatics, Göttingen; [2] University Medical Center Göttingen, Central Biobank UMG, Göttingen; [3] University of Göttingen, Campus-Institute Data Science (CIDAS), Göttingen; [4] University of Göttingen, Cluster of Excellence "Multiscale Bioimaging: from Molecular Machines to Networks of Excitable Cells" (MBExC), Göttingen

Introduction

In biomedical research, laboratory material management is crucial for FAIR documentation and lifecycle tracking of resources, enabling auditability and reproducibility. However, the requirements often surpass the limitations of standard tools for research data management (RDM) due to the increasing specialization of materials such as induced pluripotent stem cells (iPSCs) and adeno-associated viral vectors (AAVs). Over the last decade, our group has developed menoci [1], a web portal for cross-linking and sharing research results and underlying resources within consortia. Through our experience in designing and implementing several material catalogue modules, we have encountered recurring challenges for laboratory material management and its documentation.

Methods

We collected and reviewed our experiences with the development of customized menoci catalogues for various materials to identify common challenges and user requirements in diverse practical use cases from our consortia. We critically re-analyzed our approaches and also outlined our plans how to overcome remaining challenges during ongoing iterative software development.

Results

Our use cases range from the routine management of antibodies to the more specialized documentation of patient-derived iPSCs and preclinically applied AAVs. Recurring requirements include semi-automated exports for mandatory documentation of genetically modified organisms and interfaces for upload of metadata to (inter)national repositories like hPSCreg® [2] for iPSCs. Furthermore, redundant and hence time-consuming data entry and the lack of a single source of truth when dealing with multiple systems are major concerns of researchers in all settings. Therefore, we provide selectable templates, enable cloning of datasets to serve as templates for new entries and implement synchronization approaches, which entails the extension of API functions of our own platform.

Discussion

Despite the large number of existing tools, there is still a need for customization and a lack of interoperability between platforms to combine available components into a functional RDM framework. Our experience highlights the need to include the researchers' perspective into prioritization to not only ensure effective laboratory material management but also provide sufficient practical value for users to sustain their engagement in FAIR data sharing.

[1] doi:10.1186/s12859-020-03928-1 [2] https://hpscreg.eu/

Acknowledgements: Human resources were mainly funded by DFG [CRC1565 (469281184, Z), CRC1690 (528760423, INF), Germany's Excellence Strategy - EXC 2067/1-390729940].

[28-P1-26] Design and Development of an Association Results Relational Database for Whole Genome Sequence Studies

Presenter: Andreas Ziegler

Authors: Andreas Ziegler [1, 2, 3, 4]; Linlin Guo [2, 5, 3]; Georgios Koliopanos [1]; Amra Amra Dhabalia Ashok [1]; Tanja Zeller [6, 7]; Raphael O. Betschart [1, 8]; Raphael Twerenbold [9, 5, 2]; Cristian Riccio [1]

Affiliations: [1] Cardio-CARE, Davos Wolfgang; [2] Department of Cardiology, University Heart & Vascular Center Hamburg, University Medical Center Hamburg- Eppendorf, Hamburg; [3] Center for Population Health Innovation (POINT), University Heart and Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg; [4] Discipline of Statistics, School of Agriculture and Science, University of KwaZulu-Natal, Pietermaritzburg; [5] German Center for Cardiovascular Research (DZHK), partner site Hamburg/Kiel/Lübeck, Hamburg; [6] Institute for Cardiogenetics, University of Lübeck, Lübeck; [7] German Center for Cardiovascular Research (DZHK), partner site Hamburg/Kiel/Lübeck, Lübeck; [8] Institute of Cardiogenetics, University of Lübeck, Lübeck; [9] University Heart and Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany, Hamburg

Introduction

Advances in multi-omics technologies have promoted multimodal research of complex cardiovascular conditions. Integrating and querying results across large public and in-house datasets can be challenging for diverse stakeholders.

Methods

To overcome this challenge we turned towards relational databases. We evaluated commercial solutions that could help develop a database to house the GWAS catalog, FAVOR, and association results. Product specifications were defined in close collaboration with biologists, clinicians, data scientists and bioinformaticians.

Results

This process led to the identification of key product requirements and a vendor to implement them. These requirements include data upload and maintenance, graphical user interface, identity and access management, and data protection. The resulting system enabled efficient querying and interactive visualizations of results with allowing users to easily access and explore results. Thus, a safe, scalable, informative, and user-friendly platform was deployed.

Discussion

In conclusion, our relational database offers a collaborative and multimodal exploration of cardiovascular phenotypes.

[28-P1-27] Context-driven questionnaires for medical and sports science research

Presenter: Klaus-Hendrik Wolf

Authors: Lennart Drescher [1]; Peter Düking [2]; Michael Marschollek [3]; Klaus-Hendrik Wolf [3]

Affiliations: [1] Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Braunschweig; [2] Department of Sports Science and Movement Pedagogy, Technische Universität Braunschweig, Braunschweig; [3] Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Hannover

Introduction

Continuously worn wearables become more important to answer scientific questions in medicine and sports science. While existing platforms support the collection and analysis of sensor data, the interpretation often requires further contextual information. Thus, delivering questionnaires to users based on their gathered data can be helpful.

Methods

Requirements for a data platform to manage sensor data were identified. The open-source RADAR-Base platform met most of the requirements [1]. Besides relevant server components, it encompasses a smartphone app called 'Active Remote Monitoring (aRMT)'-App enabling participants to submit scheduled questionnaires.

To enable context-driven questionnaires, the existing framework was extended with (1.) real-time event detection module, analyzing incoming sensor-data and (2.) a dynamic questionnaire scheduler that triggers surveys based on predefined events. Furthermore, the existing aRMT app had to be enhanced to support context-driven questionnaires.

Results

The extended platform allows researchers to define events (e.g., completion of a training session, first 100 steps of the day, sustained heart rate elevation). The new questionnaire scheduler processes incoming sensor streams and triggers questionnaires upon event detection. A usability evaluation showed that data from three different sensor manufacturers could be submitted using both Android and iOS smartphones. All questionnaires were delivered as intended however, minor delays were observed due to intermittent data transfer patterns.

Discussion

The exteded RADAR-Base enables context-driven questionnaires, improving the timeliness and relevance of self-reported data. This approach enhances correct interpretation of sensor data by adding relevant contextual information from users. A known limitation remains the non-real-time transfer of data from wearables to smartphones and on to RADAR-Base typically performed in bursts to conserve battery power. This influences the timely delivery of questionnaires. Nevertheless, the system demonstrates the feasibility of context-driven questionnaires for use in upcoming studies.

References

[1] Ranjan Y, Rashid Z, Stewart C, Conde P, Begale M, Verbeeck D, et al. RADAR-Base: Open Source Mobile Health Platform for Collecting, Monitoring, and Analyzing Data Using Sensors, Wearables, and Mobile Devices. JMIR MHealth UHealth. 20197:e11734. https://doi.org/10.2196/11734

[28-P1-28] Key steps in data management for a longitudinal registry on cardiac rehabilitation in the Swiss Alps: the RECOVER cohort

Presenter: Pietro Sartori

Authors: Pietro Sartori [1]; Jan Rieß [2, 3]; Gloria Petrasch [2]; David Niederseer [2, 4, 5]; Andreas Ziegler [1, 3, 6, 7]

Affiliations: [1] Cardio-CARE, Medizincampus Davos, Davos; [2] Hochgebirgsklinik Davos, Medicine Campus Davos, Davos; [3] Department of Cardiology, University Heart & Vascular Center Hamburg, University Medical Center Hamburg- Eppendorf, Hamburg; [4] Christine Kühne-Center for Allergy Research and Education (CK-CARE), Medicine Campus Davos, Davos; [5] Center of Translational and Experimental Cardiology (CTEC), Department of Cardiology, University of Zurich, Zurich; [6] Center for Population Health Innovation (POINT), University Heart and Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg; [7] Discipline of Statistics, School of Agriculture and Science, University of KwaZulu-Natal, Pietermaritzburg

Introduction

Modern cohort studies collect increasingly large amounts of data to perform countless analyses. It is thus important to clearly define the data to be collected, establish appropriate storage methods, and analytical techniques to ensure the study produces valid and meaningful results. These key points were followed in RECOVER (Cardiac REhabilitation COhort at the Medicine Campus DaVos to invEstigate Recovery). The aim of this work is to demonstrate how a digital data infrastructure can be successfully developed and implemented within a cohort study.

Methods

The creation of the database proceeded in three specific phases. The first phase consisted in defining the variables to be collected. Physicians and researchers, in collaboration with biostatisticians and data scientists determined the variables and patient reported outcomes to be recorded. The variables were defined in an MS Excel codebook and consisted in multiple worksheets which were organized by topic. The second phase involved coding the script in R and converting it into a REDCap-compatible data dictionary. The most important aspects of the code were to keep variable names as simple as possible and to translate validation, ranges, and the branching logic into the data dictionary for compatibility with REDCap. The third phase involved uploading the data dictionary to REDCap and implementing a multi-stage quality control process with feedback loops to the previous step.

Results

More than 4000 variables across 61 instruments were entered into REDCap, across 61 instruments, covering 5 timepoints. Validity, range, and branching logic checks for each variable were performed at various stages, including the research phase, the coding phase, and the implementation phase of the study. In addition, 23 original questionnaires, including 59 patient reported outcomes were redesigned in REDCap for complete digitalization to reflect the originals as closely as possible, in accordance with the guidelines for each questionnaire.

Discussion

The process of collecting variables in MS Excel, converting them into a data dictionary and automatically integrate them with an R script into REDCap is feasible and adjustable at any stage. Due to its flexibility, this data management system is scalable and offers the flexibiility of expansion to other studies.

[28-P1-29] CASPA: A repository for residential histories to enable environmental health analysis in the NAKO health study

Presenter: Alexander Koglin

Authors: Alexander Koglin [1]; Robert Wolff [1]; Kathrin Wolf [2]; Marco Dallavalle [2]; Wolfgang Hoffmann [3]; Gunthard Stübs [3]

Affiliations: [1] Trusted Third Party of the University Medicine Greifswald, Greifswald, Germany, Greifswald; [2] Helmholtz Zentrum München - Deutsches Forschungszentrum für Gesundheit und Umwelt (GmbH), Munich; [3] Universitätsmedizin Greifswald KöR, Institut für Community Medicine, Greifswald

Introduction

The NAKO Health Study is a large long-term cohort study, which examines the health of a representative sample of the population. Yet precise life course-related residential histories, essential for investigating e.g., the influence of environmental factors (including pollution, noise, heat/climate conditions) and regional characteristics (e.g., deprivation, medical care) on human health. With CASPA, we aimed to develop a system that serves as the leading repository for historical residential data for the NAKO.

Methods

Heterogeneous address information is available from several sources such as recruitment, direct contacts, population registers, retrospective surveys and vital status follow-up [1]. We set up a multi-stage process to combine the fragmented address information into continuous quality assured, geocoded residential histories:

Data integration: Using address events from various sources near real-time Address matching: Automated linkage of addresses to generate unique geocodes Aggregation: Combining addresses into subsequent residential episodes Quality assurance: Manual resolution of conflicts and data gaps

by study staff For the linkage with environmental exposures, identities, addresses, and geocodes are strictly separated. Researchers are provided only with the minimum sufficient information with pseudonymized, time-aggregated exposure variables to ensure the highest data protection standards.

Results

Around 50% of the participants moved in a timeframe of 5 years and more than 1 million address entries were available since the NAKO baseline examination. We currently achieve a nearly complete, failure tolerant matching quality across all participants. Finally, the automated aggregation of address data, combined with the possibility of manual corrections, guarantees high data validity for long-term epidemiological analyses.

Discussion

CASPA provides a tool that integrates isolated, fragmented and heterogeneous address information into continuous residential episodes enabling linkage with geographical information, e.g. analysing health effects of environmental exposures. By combining a rule-based, reproducible aggregation algorithm with a systematic priorization of sources (e.g., population registers vs. self-reported information), a high level of data quality is ensured.

[1] Greiser KH, Bohn B, Hilger-Kolb J, Panreck L, Breunig E, Lieb W, Pischon T, Schikowski T, Völzke H. The German National Cohort (NAKO): Design, current state, and further follow-up data collection. Eur J Public Health. 2023 Oct 2433(Suppl 2):ckad160.499. doi: 10.1093/eurpub/ckad160.499. PMCID: PMC10596689

[28-P1-30] Integration of Result Datasets from Use and Access in the German Na-tional Cohort (NAKO) - Workflows, Tools and Achievements

Presenter: Gunthard Stübs

Authors: Jacob Bräutigam [1]; Henry Mehlan [1]; Wolfgang Hoffmann [1]; Gunthard Stübs [1]

Affiliations: [1] Universitätsmedizin Greifswald KöR, Institut für Community Medicine, Greifswald

Introduction

The German National Cohort (NAKO) is the largest population-based health study in Germany. It involved data collection from 205,000 participants. In addition to primary data, the dataset is supplemented by secondary data from external sources. The results data provided by researchers through the Use and Access procedure are a third important source of data. Here we describe the NAKO process for the integration of results data.

Methods

The requirements and process analysis revealed that the first step is to integrate the descriptive metadata into the NAKO data dictionary. This data must be defined by the researchers without any technical barriers. The storage structures in a database must be generated from this metadata. The

data integration software must first and foremost resolve the project-specific pseudonyms. However, it must also verify the current status of consent and the plausibility of the data itself. After data integration, the new variables must be displayed in the data directory and be selectable for other researchers.

Results

The designed and implemented process starts with the definition of the additional variables in a simple Excel table. This can be converted into the data dictionary and a relational table is generated in the study database. We developed a web-technology based data integration management system (DIMAS). DIMAS facilitates the data integration and comprises the re-pseudonymization, consent verification, completeness and data type checks as well as a few correction methods. The data integration triggers the display of the additional variables in the data directory of our application management web-application TransferHub. Heretofore, we have performed this process a several dozen times. The data integrations resulted in a few thousand additional data elements now supplementing the NAKO dataset by a two-digit percentage.

Discussion

With this work we streamlined and standardised the NAKO results data integration process. With DIMAS we developed a specialized ETL tool. However, to avoid the dilution of the NAKO data dictionary, it is important to evaluate carefully in advance which results datasets to integrate and which not. The most significant data are those from limited resources, like biosamples, or with costly generation, like MRI. We have integrated and redistributed such datasets successfully.

[28-P1-31] Better Research - Data Extraction: A Cross-Context Guide

Presenter: Dominik de Sordi

Authors: Antje Timmer [1]; Gesine Seeber [1]; Rainer Röhrig [2]

Affiliations: [1] Carl von Ossietzky Universität Oldenburg, Oldenburg; [2] Universitätsklinik RWTH Aaachen, Aachen

Introduction: Many types of medical research use data extraction to obtain information suitable for analysis. Although the quality of the extracted data is always a prerequisite for reliable research results, the importance attributed to this phase varies considerably across research contexts. Methods evaluation results and research standards, as available for data extraction in systematic reviews (SRs), are only partially accessible to researchers in other fields. This work aims to provide practical guidance for beginners in data extraction across different research contexts.

Methods: The work is primarily based on the authors' long-standing experience and synthesises insights from various research fields. Additionally, guidelines, reviews, and study results from the literature on data extraction in SRs, automation processes, the use of electronic patient data for research purposes, extraction software, and related topics served as a foundation. The focus is on practical applications that could also be implemented in a manual "paper and pencil" format.

Results/Implementation: We situate the data extraction phase within the context of a research project and the growing possibilities for automation. Step by step, the user is then guided through the phases of form creation, testing, and application using practical examples. We address the concepts of information completeness, efficient and user-friendly data entry, information accuracy and reproducibility, data protection compliance, and compatibility with statistical programming requirements for analysis. Where possible, we refer to technical solutions, further reading, and relevant current developments.

Discussion: Despite - or perhaps because of - the increasing availability of digitised, structured clinical data and automation options, practical knowledge of data extraction remains a relevant, yet often neglected, component of good research practice. With this guide, we are closing a training gap.

PS - We will shortly submit this guide to egms MIBE the full paper needed a few more days beyond the deadline.

P2 - Prediction and Personalized Medicine

Poster area: P2

[28-P2-01] ISO Recommendations and Requirements for Predictive Computational Models in Personalised Medicine

Presenter: Martin Golebiewski

Authors: Martin Golebiewski [1]; Laura Lopez-Perez [2]; Elena Martinelli [3]; Tito Poli [3]

Affiliations: [1] HITS gGmbH, Heidelberg; [2] Universidad Politécnica de Madrid, Madrid; [3] Department of Medicine and Surgery, University Hospital of Parma, Parma

Introduction

Health data are used to set up computational models for simulations and predictions, e.g. for in silico clinical trials, disease prognosis, treatment predictions or patient stratification. These data need to be interoperable for model integration or validation. Thus, standards are crucial for structuring, describing, and associating models and data, as well as for describing how components interact, are linked and embedded in their context.

Methods

The ISO committee ISO/TC 276/WG 5 Data Processing and Integration has released standards that provide recommendations and requirements for in silico modelling in personalized medicine.

Results

ISO 20691:2022 provides a framework for the development and application of interoperable (meta-)data standards in the life sciences. It defines requirements and rules for the application and design of standards for formatting, description, and documentation of data.

The standard series ISO 9491 provides a guideline with recommendations and requirements for predictive computational models in personalized medicine research. It has two parts, the first ISO 9491-1 'Guidelines for constructing, verifying, and validating models' has been published as ISO technical specification and recently revised. It gives recommendations and requirements for data preparation, integrating data into models, model formatting, validation, simulation, storing, sharing and application. The annex lists specific community standards and widely used modelling approaches for in silico approaches in personalized medicine. The second part ISO/TS 9491-2 provides guidelines for implementing computational models in clinical integrated decision support systems and contains recommendations for computational models and their application in the clinical research context.

To complement these activities, a new joint working group is currently formed between the ISO/TC 276 Biotechnology and IEC/TC 62, the IEC technical committee developing standards for medical equipment, software, and systems, to address the standardization of the credibility of computational modelling used in medical devices. The group addresses the verification, validation, and uncertainty quantification of models. This work will build on the national US standard ASME V&V 40:2018 for assessing the quality and model credibility of medical devices and extend it to an international standard.

Discussion

These ISO standards will support the modelling and data integration processes in personalized medicine, rendering the models and their results more reliable.

[28-P2-02] Developing a Classification System for Juvenile Idiopathic Arthritis Using Latent Class Modeling

Presenter: Giulia Capitoli

Authors: Alessandra Alongi [1]; Giulia Capitoli [2]; Saverio La Bella [1]; Alberto Martini [3]; Nicolino Ruperto [1]; Stefania Galimberti [2]

Affiliations: [1] School of Medicine and Surgery, University of Milano-Bicocca, Monza; [2] Bicocca Bioinformatics Biostatistics and Bioimaging Centre - B4, School of Medicine and Surgery, University of Milano-Bicocca, Monza; [3] University of Genova, Genoa

Introduction

Juvenile Idiopathic Arthritis (JIA) encompasses clinically heterogeneous entities with overlapping features that complicate classification and clinical decision-making. Latent class models (LCM) are mixture models that identify homogeneous subgroups. To translate unsupervised clustering into a usable clinical tool, we employed a multi-step pipeline combining development and validation of cluster-based classification rules with expert consensus.

Methods

We analyzed a prospective multinational cohort of 1,509 patients with new-onset JIA, including clinical and laboratory variables selected for clinical relevance by experts in a prior Delphi exercise. The cohort was split into a training set (n = 857) and an independent validation set (n = 652) from distinct centres.

In Step 1, missing data were imputed using Random Forest LCM was fitted on the training set. The optimal number of classes was selected based on information criteria and clinical interpretability.

In Step 2, we used multinomial elastic-net regression to select a reduce pool of predictors discriminating the identified latent classes. An exhaustive model selection procedure using class-specific logistic regression evaluated all predictor combinations and ranked models by corrected Akaike Information Criterion (AICc) to identify parsimonious set of predictors.

In Step 3, scores were derived by rounding model coefficients, and optimal cut-offs were defined using the Youden index.

In Step 4, LCM was refitted on the validation set using the selected k, and the performance of the scores in predicting class membership was assessed. The top ten classification rules for each class based on AICc were proposed to clinical experts for evaluation in an ongoing Delphi study.

Results

Six latent classes with distinct characteristics were identified. The LCM refitted in the validation set showed similar class-conditional feature distributions, supporting the selected solution. The classification system based on the top-ranking models for each class achieved a Macro-F1 score of 0.80 (95% CI: 0.747-0.840) in predicting class membership on the validation set. Consensus expert evaluation is ongoing.

Discussion

We derived a clinically interpretable JIA classification system from unsupervised clustering by combining LCM, exhaustive model exploration, and expert consensus, identifying parsimonious scoring systems with good discrimination in the validation set. This integrated approach may support development of clinically meaningful classification tools in complex diseases.

[28-P2-03] A scoping review of studies updating a clinical prediction model

Presenter: Amardeep Legha

Authors: Amardeep Legha [1]; Joie Ensor [1]; Kym Snell [1]; Lucinda Archer [1]; Nicola Adderley [1]; Joseph Alderman [1]; Glen Martin [2]; Karel Moons [3]; Paula Dhiman [4]; Maarten van Smeden [5]; Alastair Denniston [1]; Gary Collins [1]; Richard Riley [1]; Rebecca Whittle [1]

Affiliations: [1] University of Birmingham, Birmingham; [2] University of Manchester, Manchester; [3] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht; [4] University of Oxford, Oxford; [5] University Medical Center Utrecht, Utrecht

Introduction

Clinical prediction model performance may deteriorate over time or when applied in new settings, limiting transportability. Hence, to improve model performance, it may be important to update an existing model. However, there is limited evidence on what triggers model updating, how models are updated, and whether sample sizes for updating are sufficient and adequately justified. We undertook a scoping review to identify studies that update clinical prediction models to examine the reported rationale, methods used, sample size justification, and whether any reporting guidelines were adhered to.

Methods

We searched MEDLINE (via OVID) for primary studies published in 2025. Eligible studies were external validation studies of regression-based clinical prediction models that included model updating using any approach. Screening was conducted by a single reviewer, with a second reviewer consulted for ambiguous records. Findings were summarised using counts and percentages.

Results

Thirty-three studies reporting model updating were identified. A range of updating approaches were used, including updating the intercept only, updating both the intercept and slope, and revising model coefficients.

Reporting of the rationale for model updating was limited: in 7 studies (21%) no rationale was provided 7 (21%) aimed to assess the impact of adding new predictor(s) 3 (9%) cited challenges obtaining predictors in the developed model and 16 (48%) aimed to improve external validation performance.

No studies reported performing a sample size calculation specifically for model updating. Sample size calculations were reported for external validation in 10 studies (30%), five stated that their sample size was sufficient without providing a calculation, and no justification for the sample size used was given in the remaining 18 studies (55%).

Eleven studies (33%) did not state adherence to any reporting guidelines. Among the 22 studies that did, TRIPOD or TRIPOD+AI were most commonly used (n=19).

Discussion

Studies updating clinical prediction models often provide limited rationale for undertaking model updating and rarely justify the adequacy of the sample size used. Better reporting of updating rationale, and development of guidance for determining sample size required for model updating, is needed.

[28-P2-04] Sample size requirements for updating prediction models with a binary outcome: an evaluation of existing criteria

Presenter: Rebecca Whittle

Authors: Rebecca Whittle [1]; Richard Riley [1]; Gary Collins [1]; Kym Snell [1]; Amardeep Legha [1]; Lucinda Archer [1]; Joie Ensor [1]

Affiliations: [1] University of Birmingham, Birmingham

Introduction

Many clinical prediction models need updating over time to remain reliable, for example due to shifts in outcome prevalence or case-mix. However, there is limited guidance on sample size requirements for studies updating a model. In practice, researchers often use criteria proposed for model development though their suitability for updating is unclear. We aimed to evaluate this for studies updating prediction models for a binary outcome using logistic regression.

Methods

We conducted a simulation study examining two strategies: intercept-only updating (shift in outcome prevalence) and updating of both intercept and calibration slope (shift in predictor effects). Development datasets (N=100,000) were generated with 12 predictors, c-statistics ranging from 0.6 to 0.9, and outcome prevalences 0.1 to 0.5. Updating populations (N=100,000) were generated with shifted outcome prevalences for slope updating scenarios, predictor effects were also modified. Recommended minimum sample sizes were obtained using existing closed-form calculations of Riley et al. [1], targeting precision of the overall outcome proportion, an expected calibration slope of 0.9, and expected optimism of 0.05 in R2. For each scenario and sample size criterion, 500 samples were drawn and models refitted. Performance was evaluated by precision of the estimated outcome proportion in the updating sample, the calibration slope, and optimism in R2.

Results

For intercept-only updating, Riley sample size calculations achieved the target for the mean precision of the overall outcome proportion, for both fixed (mean margin-of-error$\leq$0.05) and adaptive (based on the outcome prevalence) precision targets, though variability across datasets was substantial.

For intercept and slope updating, the target mean optimism in R2 was achieved in most simulations, but the calibration slope target was met less consistently, and also showed large variability. To address this, we demonstrate how simulation-based approaches [2,3] can instead identify sample sizes that achieve both target mean performance and high assurance in a single dataset.

Discussion

Existing sample size criteria for model development may not ensure that desired performance targets are achieved with high probability in model updating settings. We recommend using a simulation-based approach instead, to target assurance-based sample sizes.

[1]doi:10.1136/bmj.m441

[2]doi:10.1186/s12874-024-02268-5

[3]doi:10.48550/arXiv.2504.18730

[28-P2-05] Handling of data limitations in clinical prediction models developed using electronic health records: A focused scoping review

Presenter: Rebecca Whittle

Authors: Rebecca Whittle [1]; Kym Snell [1]; Joie Ensor [1]

Affiliations: [1] University of Birmingham, Birmingham

Introduction

Clinical prediction models developed using routinely collected electronic health record (EHR) data are increasingly used to support diagnosis, prognosis, and treatment decisions. However, EHR data are affected by selective measurement, missingness, coding inaccuracies, imprecise timing, and incomplete outcome recording, all of which may introduce uncertainty into predictions and clinical decisions. It is unclear how often recent EHR-based regression model development studies address these issues explicitly. This study aimed to characterise how such studies handle these data limitations.

Methods

We conducted a focused scoping review of recent regression-based clinical prediction model development studies using routinely collected EHR data. MEDLINE was searched for studies published in 2025. Screening was undertaken in reverse chronological order to obtain a recent sample of studies. For each included study, we extracted information on baseline definition, predictor measurement window, handling of missing predictors, outcome ascertainment, acknowledgement of EHR-related limitations, and use of sensitivity analyses to address these limitations.

Results

Thirty studies were included across a range of clinical areas and settings. Twenty of 30 studies (67%) clearly defined a baseline timepoint. In 12/30 (40%), the predictor measurement window was not clearly defined, while the remainder typically relied on look-back periods of variable length. In the 22/30 (73%) studies that reported their approach to handling missing predictor data, methods included complete-case analysis (6/22, 27%), multiple imputation (6/22, 27%), single imputation (5/22, 23%), use of missing data categories (3/22, 14%), and assumption-based approaches (2/22, 9%). Twenty of 30 studies (67%) acknowledged at least one EHR-related limitation, most commonly missing data, coding inaccuracies, and potential misclassification. However, only 8/30 studies (27%) reported any sensitivity analysis directly relevant to EHR-related assumptions, such as removing predictors subject to strong assumptions and using alternative missing data strategies.

Discussion

EHR-based regression prediction model studies rarely evaluate how assumptions about EHR data influence predicted risks or clinical decision-making. As a result, the extent to which these assumptions introduce uncertainty into model outputs and clinical decisions is largely unknown. This highlights a need for methodological work to quantify this uncertainty and for strategies to mitigate its impact in practice.

[28-P2-06] An evaluation of variable selection methods for developing binary risk prediction models at recommended sample sizes

Presenter: Xinrui Su

Authors: Xinrui Su [1]; Gareth Ambler [1]; Nathan Green [1]; Menelaos Pavlou [1]

Affiliations: [1] University College London, London

Introduction

Clinical prediction models are developed by modelling the association between predictor variables and a health outcome. Sample size calculations for developing such models typically assume a pre-specified set of predictors, and that variable selection will not be used. However, variable selection methods are almost always used in practice, sometimes with the intention of 'meeting' sample size assumptions when the sample size is too small to support all candidate predictors. This scenario has received relatively little attention in the literature.

Methods

We conducted a simulation study to evaluate the predictive performance of logistic regression models developed using univariable selection and backward elimination (with various inclusion thresholds), LASSO (with various tuning methods) and commonly used combinations thereof. The full model fitted by maximum likelihood estimation served as the reference. We considered scenarios with varying relative strengths of predictors and assessed each method at the recommended sample size, as well as smaller and larger sample sizes. Our aim was to explore whether the use of variable selection allows relaxation of sample size requirements, or whether it may in fact necessitate larger sample sizes.

Results

We found that LASSO with 1-SE tuning showed inferior predictive performance in all scenarios. In contrast, LASSO with modified tuning (LASSO-mod) generally performed well in all scenarios. The performance of univariable selection, backwards elimination and LASSO with standard tuning was more dependent on the specific scenario.

Discussion

We recommend following existing guidance to calculate the required sample size based on the number of all candidate predictors when developing a clinical prediction model, then using LASSO-mod to estimate the model. When the available sample size is lower than that recommended, using variable selection to select a smaller number of predictors based on which the sample size requirements are met, is not advisable. Instead, use of LS-mod is likely to result in models that are better than the full model.

[28-P2-07] Cost-based variable selection to improve clinical utility of prediction models

Presenter: Lasai Barreñada

Authors: Lasai Barreñada [1, 2]; Andrew Vickers [3]; Ewout Steyerberg [4]; Dirk Timmerman [1]; Laure Wynants [5, 2]; Ben Van Calster [1, 2]

Affiliations: [1] Department of Development and Regeneration, KU Leuven, Leuven; [2] Leuven Unit for Health Technology Assessment Research (LUHTAR), KU Leuven, Leuven; [3] Memorial Sloan Kettering Cancer Center, New York; [4] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht; [5] Maastricht University, Maastricht

Introduction

Clinical prediction models guide critical diagnostic and prognostic decision-making. However, traditional variable selection methods rely primarily on statistical properties, frequently ignoring clinical utility and measurement costs. Consequently, this results in models with predictors that are too expensive for routine clinical practice. This study introduces a novel variable selection algorithm that chooses predictors by maximizing clinical utility while explicitly accounting for measurement costs.

Methods

Clinical utility is measured by Net Benefit (NB), which balances true positives against weighted false positives. We propose a variable selection algorithm that adjusts standard NB to account for predictor measurement costs. The algorithm evaluates all possible predictor combinations, incorporates their costs, and selects the model with the highest cross-validated NB. For high-dimensional settings, we introduce an NB-based backward elimination alternative. Our algorithm also calculates variable importance based on each predictor's contribution to clinical utility. We illustrate this algorithm through a simulation and validate it in an ovarian cancer diagnostic case study, comparing cost-adjusted and unadjusted NB methods against LASSO and standard backward elimination. The algorithm and variable selection visualization tools are available in the NBvarsel R package (https://lasaibarrenada.github.io/NB_varsel/)

Results

Our results show that the proposed algorithm correctly identifies important predictors and ranks predictor combinations based on the model's NB. Crucially, optimal variable selection during model development translates directly to optimal external performance. While all selection procedures perform comparably when measurement costs are ignored, accounting for costs allows our algorithm to select a simpler model. This cost-adjusted model achieves superior clinical utility with minimal loss in discrimination and calibration.

Discussion

Using NB for variable selection allows us to account for the cost of measuring predictors in terms of how many patients we are willing to test to detect an event. This approach increases the overall utility of the model by removing costly variables that do not meaningfully improve predictions, maintaining model performance while reducing clinical burden.

[28-P2-08] Impact of feature correlation and missing data structure on predictive model performance

Presenter: Sergio Sabroso-Lasa

Authors: Sergio Sabroso-Lasa [1]; Luis Mariano Esteban [2]; José Tomás Alcalá [3]; Mireia Andueza [1]; Núria Malats [1]

Affiliations: [1] Spanish National Cancer Research Centre (CNIO), Madrid; [2] Escuela Universitaria Politécnica de La Almunia, Universidad de Zaragoza., Zaragoza; [3] Universidad de Zaragoza, Zaragoza

Introduction

The exponential growth of large, complex, multi-source datasets has made handling missing data a critical challenge in statistical modelling, since the percentage and distribution of missing data can significantly affect analytical outcomes. Effectively imputing missing values is essential to ensure the reliability of statistical inference and predictive modeling. Although many imputation methods exist and prior studies have assessed the impact on model performance of the proportion of missing data depending on the imputation model employed, the effect of the underlying correlation structure among features with missing values remains a complex issue that requires further investigation.

Methods

We conducted simulations under realistic conditions to assess the impact of the imputation method used, the proportion and distribution of missing data, and the correlation structure on performance metrics such as AUC, sensitivity, specificity, and the Youden index. These scenarios included normally distributed variables with varying predictive capacities and across different partial correlation structures ranging from -0.8 to 0.8. In these partially correlated scenarios, we compared results obtained with the MissForest and Multivariate Imputation by Chained Equations (MICE) methods across missingness levels ranging from 5% to 90%. All results are presented using interpretable scatter plots that illustrate, according to the initial correlation, the changes in performance metrics and the correlation loss as the proportion of imputed data increases.

Results

When missing data are present, diagnostic performance decreased by 20-30% compared to complete-data models. This decrease is less pronounced under positive correlations and more pronounced when the variables are positively correlated, with independence falling in between. This pattern appears driven by a loss of the underlying correlation structure, particularly pronounced when initial correlations are negative and more evident with MissForest than with MICE.

Discussion

The results highlight the significant impact of the structure of missing data on various diagnostic metrics. Although most modern predictive models incorporate an imputation phase, most do not examine how the distribution of missing data influences model performance. This oversight prevents

potential improvements in evaluation metrics and leads to unreliable Youden Index values and cutoff points, underscoring the difficulty of defining optimal clinical cutoffs in incomplete datasets.

[28-P2-09] Evaluating the Impact of Imputation Under Missing-at-Random and Missing-Not-at-Random Mechanism on the Performance of Treatment Response Prediction Models

Presenter: Mian Haider Ali

Authors: Mian Haider Ali [1]; Ulrich Mansmann [2]

Affiliations: [1] Ludwig Maximilian University of Munich, Munchen; [2] Universität München, München

Introduction

missing outcome data are common in longitudinal clinical and epidemiological studies and can influence the development and evaluation of prediction models (PM). Many standard imputation methods assume that data are missing at random (MAR), but this assumption may be unrealistic when dropout or non-response depends on the unobserved outcome itself, that is, missing not at random (MNAR). In such settings, the imputation strategy may affect not only parameter estimates, but also model performance, calibration, and generalizability. This study examines how MAR- and MNAR-based imputation methods influence treatment response prediction for continuous longitudinal outcomes.

Methods

we use a simulation framework with repeated outcome measurements at baseline and at least four follow-up time points, together with subject-level covariates and randomized treatment assignment. Complete data are generated from a prespecified longitudinal data-generating mechanism. Missingness is then introduced at the final follow-up under both MAR and MNAR scenarios, with different levels and structures of missingness. Under MAR, missingness depends on observed baseline severity, whereas under MNAR it depends on the unobserved final outcome in practice, both mechanisms may occur together. In the training phase of the PM, missing information is imputed under both mechanisms. Particular attention is given to cluster-informed imputation strategies motivated by heterogeneity in outcome trajectories. Individuals are grouped into response-profile clusters using partitioning around medoids (PAM) with Gower distance, which allows clustering in the presence of missing longitudinal outcomes. These clusters are then incorporated into a regression-based imputation model with cluster-specific shifts within a pattern-mixture framework. The proposed method will also be applied to attention deficit hyperactivity disorder (ADHD) clinical trial data.

Results

analyses compare the effect of different imputation assumptions on PM quality including prediction error and calibration, across repeated simulations.

Discussion

this study evaluate whether misspecification of the missing-data mechanism reduces predictive performance. The findings are expected to provide practical guidance on the use of MAR and MNAR imputation strategies when developing treatment response PMs from incomplete longitudinal data.

[28-P2-10] Real-Time Handling of Missing Predictors in Clinical Prediction Models: A Simulation Study and Maternal Electronic Health Record Case Study

Presenter: Mae Chester-Jones

Authors: Mae Chester-Jones [1]; Gary Collins [2]; Stephen Gerry [3]

Affiliations: [1] University Of Oxford, Oxford; [2] Department of Health Sciences, School of Health Sciences, college of Medicine and Health, University of Birmingham, Birmingham, United Kingdom, Birmingham; [3] University of Oxford, Oxford

Introduction

Clinical prediction models integrated within electronic health records must provide real-time predictions and requires strategies for handling missing predictors at the point of use (implementation). The method for handling missing data at implementation should be evaluated at validation, and consistent across the clinical prediction model pipeline.

Methods

We compared four strategies for handling incomplete predictor variables in real-time: regression imputation, reference values, pattern sub-models and XGBoost. Regression imputation and reference values impute the missing predictor with the average value (expected value from a set of covariates and subgroup means respectively), whereas the pattern sub-model fits separate models for each observed missingness pattern and the XGBoost is a tree-based machine learning model that learns the optimal decision tree for missing data without requiring imputation.

Scenarios were simulated for development and validation with a range of missing data mechanisms (MCAR, MAR, MNAR), proportions of missing data (25%, 50%, 75%) and sample sizes (N=500, N=10,000, N=100,000). The strategies were then applied to four years of maternal electronic health record data from two University Hospitals Birmingham NHS Foundation Trust sites.

Results

These strategies demonstrated identical performance across mechanisms in simulated scenarios with 25% of data missing. When the underlying mechanism for missingness and for the outcome depended on the same variables, the pattern sub-model and XGBoost experienced miscalibration when the proportion of missingness at validation (and implementation) differed from that of the derived models. Similarly to the simulation study, the discriminative performance was similar at validation with evidence of miscalibration of the pattern sub-model and XGBoost slopes.

Discussion

These strategies demonstrated identical performance across mechanisms in simulated scenarios with 25% of data missing. When the underlying mechanism for missingness and for the outcome depended on the same variables, the pattern sub-model and XGBoost experienced miscalibration when the proportion of missingness at validation (and implementation) differed from that of the derived models. Similarly to the simulation study, the discriminative performance was similar at validation with evidence of miscalibration of the pattern sub-model and XGBoost slopes.

[28-P2-11] Combining Multiple Imputation and Internal Model validation to Estimate Clinical Prediction Model Performance

Presenter: Sinclair Awounvo

Authors: Sinclair Awounvo [1]; Meinhard Kieser [1]; Manuel Feißt [1]

Affiliations: [1] Institute of Medical Biometry, University of Heidelberg, Heidelberg

Introduction

Accurate estimation of model predictive performance is a crucial step in clinical prediction modeling (CPM). In practice, model performance is often assessed through internal model validation (IMV) because an external dataset is frequently unavailable. When datasets contain missing values, appropriate handling of missing data is essential to obtain unbiased results. Multiple imputation (MI) has become increasingly popular for handling data assumed to be missing at random (MAR). However, this raises the question of which strategy for combining MI and IMV yields optimal results in terms of accuracy of model performance estimation, methodological complexity, and computational cost.

Methods

An extensive simulation study compares two strategies for combining MI with IMV that differ in the timing of MI. In the first strategy, MI is performed prior to IMV (MI-prior-IMV): the incomplete dataset is multiply imputed, predictions are obtained via cross-validation (CV) for each imputed dataset, and model performance is estimated from pooled predictions across imputations. In the second strategy, MI is performed during IMV (MI-during-IMV): at each CV step, the incomplete training and test datasets are multiply imputed, predictions are pooled across imputed test datasets, and model performance is evaluated from pooled predictions at the end of CV.

Both strategies are evaluated across a wide range of scenarios for binary and continuous outcomes and different modeling algorithms (regression, Random Forest, and XGBoost). Simulation scenarios vary with respect to true model performance, sample size, event proportion, number of predictors, and proportion of missing values. Several performance measures, including the accuracy of model performance estimation and computational runtime, are summarized across simulation repetitions.

Results

Preliminary results indicate that MI-during-IMV yields more accurate estimates of model performance than MI-prior-IMV, but at a substantially higher computational cost as MI is repeated within each CV iteration. This difference in computational cost is practically relevant, as it increases drastically with the number of imputations and CV folds.

Discussion

The simulation study aims to provide evidence-based recommendations on how to optimally combine MI with IMV for accurate model performance estimation, assessing whether the improved accuracy of MI-during-IMV justifies its greater complexity and computational cost.

[28-P2-12] A Framework for Assessing Imputation Methods for Survival Prediction Models via Interpretability Distortion

Presenter: Imad El Badisy

Authors: Imad El Badisy [1]; Roch Giorgi [2]

Affiliations: [1] Centre Mohammed VI de la Recherche et de l'Innovation, Rabat; [2] Aix Marseille Univ, APHM, Inserm, IRD, SESSTIM, ISSPAM, Hop Sainte Marguerite, BioSTIC, Marseille

Introduction

Missing-data imputation methods are usually evaluated using distributional checks or predictive performance. In analyses based on survival prediction models, however, these criteria do not directly assess whether imputation preserves the interpretation of the fitted model. We propose an evaluation framework based on interpretability distortion, defined as the discrepancy between interpretability summaries obtained after imputation and those obtained from a complete-data reference.

Methods

We illustrate the framework using simulated right-censored data under MCAR, MAR, and MNAR mechanisms, with overall missingness rates of 10%, 30%, and 50%, including a near-reference low-missingness scenario and settings with substantial information loss. We compare naive, kNN, CART, missForest, FAMD, missRanger, MICE, miceRF, XGBoost-based, and complete-case approaches across a panel of survival prediction models, including Cox, Aalen, elastic-net, flexsurv, random survival forest, survival deep neural network, and survival metalearner models. Distortion is quantified for partial dependence plots, time-dependent SHAP summaries, permutation variable importance, calibration curves, and Brier curves, each relative to the complete-data reference. Standard predictive performance measures are compared with interpretability distortion to assess whether good prediction is accompanied by stable interpretation.

Results

Benchmark results indicate that the proposed framework distinguishes between imputation methods beyond predictive performance alone. In the settings considered, multiple imputation and several machine-learning-based imputation methods showed lower interpretability distortion than single-imputation approaches. Some methods preserved variable rankings while altering effect sizes, indicating that acceptable predictive performance does not necessarily imply stable interpretation relative to the complete-data reference.

Discussion

Interpretability distortion provides a complementary criterion for choosing among imputation methods when the subsequent analysis relies on survival prediction models. In addition to predictive performance, the framework evaluates whether model interpretation remains stable after imputation. This perspective is relevant in applied settings, where unstable interpretations may affect scientific conclusions, and it also makes it possible to study how distortion evolves as the proportion of missing data increases and whether that evolution depends on the missing-data mechanism.

[28-P2-13] Comparison of Statistical Methods for Addressing Values Below a Lower Detection Limit

Presenter: Anna Linger

Authors: Anna Linger [1]; Carolin Herrmann [1]

Affiliations: [1] Heinrich-Heine-Universität Düsseldorf, Düsseldorf

Introduction

Omics data are often a challenging data type when it comes to statistical analysis, especially due to their high number of compounds and frequent missing values. Missing values are usually stored with a zero-value by default. However, they can be missing due to a biological reason, where the compound is truly absent, or a technical reason, where the true compound's value lies below a lower detection limit. Hence, by their representation they are non-distinguishable. Naturally, ignoring those zero values leads to biased results. For a two-group comparison, there already exists a vast selection of statistical approaches that address those challenges. However, they all have their shortcomings and have not been comprehensively compared.

Methods

We consider standard methods like one-part tests, two-part tests and mixture models, but also a selection of more recent models that either extend or improve the standard methods (e. g. DASEV or a multi-part strategy). We conducted a comprehensive simulation study with regard to type I error rate, power and bias of the log-fold change estimation. Therefore, we considered different scenarios by varying the proportions of the biological and technical missing values, the log-fold change, and the underlying distribution of the non-missing values.

Results

Our simulation study with an underlying normal distribution confirmed results of previous simulation studies regarding the standard methods: One-part tests work well for scenarios with only technical missing values and two-part tests work well for scenarios with only biological ones, while mixture models perform well for scenarios with both types of missing values being present. The more recent models show an improvement compared to the standard methods, especially in regard to bias of the log-fold change in scenarios with many missing values, while maintaining type I error rate. However, all the methods have a deficiency in power. Simulation results for non-normal distributions are currently in process.

Discussion

More recent methods seem to have the advantage of a lower bias of the log-fold change combined with higher power values, especially for scenarios with many missing values. Although improving power, there is still potential in the development for new more powerful statistical methods.

[28-P2-14] Validation Strategy Matters. Evaluating Machine Learning from Big to Small Data in Cardiovascular Risk Prediction via an Interactive R Shiny Application

Presenter: Selen Yilmaz

Authors: Ebru Turgal [1]

Affiliations: [1] Ankara University, Department of Biostatistics, Ankara

Introduction

Validation strategy and sample size can substantially influence the performance and interpretability of machine learning models in cardiovascular disease (CVD) prediction. However, their combined effects have rarely been evaluated within a unified analytical framework. This study investigated how sample size, train-test partitioning, and validation strategy affect predictive performance and stability in CVD risk classification.

Methods

A publicly available cardiovascular risk dataset containing 70,000 individuals was analyzed using Logistic Regression, LASSO Logistic Regression, and Random Forest. Models were evaluated under two train-test splits (70/30 and 80/20) and four validation strategies (5-fold cross-validation, repeated 5-fold cross-validation, 10-fold cross-validation, and repeated 10-fold cross-validation). To simulate data-constrained settings, stratified subsamples of N = 100, 300, and 500 were generated while preserving class balance. Predictive performance was assessed using area under the receiver operating characteristic curve (AUC), sensitivity, and specificity. Variable importance and feature-selection stability were also examined. An interactive Shiny application was developed.

Results

Across the full dataset, predictive performance was consistent across validation strategies. Logistic Regression and LASSO achieved the highest discrimination (AUC: 0.794-0.796), whereas Random Forest produced comparable but slightly lower AUC values (0.787-0.790). In contrast, sample size substantially influenced performance. All models showed reduced discrimination at N = 100, with performance improving as sample size increased. Random Forest demonstrated robustness under small-sample conditions and maintained stable variable-importance rankings across validation scenarios. Systolic blood pressure, age, weight, height, and diastolic blood pressure emerged as the most influential predictors. LASSO feature selection was more sensitive to sample size and train-test partitioning, with selected predictors increasing from 1-5 in smaller subsamples to 11 in the full dataset.

Discussion

Sample size influenced model performance more strongly than validation strategy. Predictive performance remained stable across validation schemes, whereas discrimination and explanatory stability were sensitive to smaller samples. Random Forest was more robust in small datasets, while Logistic Regression and LASSO achieved similar performance with greater interpretability. These findings inform the development of reliable cardiovascular risk prediction models.

The full paper pertaining to this abstract will be published in the GMDS journal "Medizinische Informatik, Biometrie und Epidemiologie (MIBE)."

[28-P2-15] When clinical prediction models do not generalize: a simulation study in liver transplantation

Presenter: Damian Brülhart

Authors: Damian Brülhart [1]; Giulia Magini [2]; Antonia Schäfer [2]; Simon Schwab [3, 1]; Ulrike Held [1]

Affiliations: [1] University of Zurich, Zurich; [2] Geneva University Hospitals, Geneva; [3] Swisstransplant, Bern

Introduction

Clinical prediction models estimate the risk of a future outcome in patients. Such models are often externally validated using independent datasets however, even when a model has been rigorously validated in a new setting and patient population, its performance across other clinical settings remains unclear. Therefore, we systematically evaluated model performance and clinical utility across diverse patient populations to quantify the limits of transportability.

Methods

Using liver transplantation as an example, we evaluated the UK donation-after-circulatory-death (DCD) risk score under varying simulated donor and recipient characteristics. Real-world data from Swiss DCD liver transplant populations were utilized to reflect realistic cohorts. The risk score's ability to predict one-year graft failure was evaluated using calibration intercept, calibration slope, area under the receiver operating characteristic (ROC) curve, and net benefit compared to the alternative strategies of transplanting all or transplanting none.

Results

The UK DCD Risk Score's performance depended heavily on the simulated population characteristics. While the score performed adequately in settings similar to those where it was derived, it was not satisfactory in others. Especially the proportion of retranplantations in the target population, which is a heavily weighted factor in the model, had a large effect on the risk score's performance.

Discussion

The study showed, using a risk score in liver transplantation as an example, that the application of a prediction model can be limited in certain external populations, and that its transportability in new settings is not guaranteed. This study highlights the importance of external validation of clinical prediction models to determine transportability to various target populations. Their application requires careful consideration and potential model re-estimation.

[28-P2-16] Incorporating Time-Dependent Discrimination into Feature Selection for Survival Prediction

Presenter: Adina Kamarudin

Authors: Adina Kamarudin [1]; Nurain Ibrahim [2]

Affiliations: [1] Universiti Malaya, Kuala Lumpur; [2] Universiti Teknologi MARA Malaysia, Shah Alam

Introduction

Feature selection is critical in the development of predictive models, particularly in high-dimensional biomedical studies. Conventional approaches typically based on static area under the receiver operating characteristic curve (AUC), evaluated at a single time point. However, in survival settings, outcomes evolve over time, and the predictive relevance of features may vary across different stages of follow-up. Although time-dependent ROC methods provide a more appropriate evaluation framework, they are typically applied only after feature selection which disconnect the model selection and evaluation.

Methods

We propose a framework that incorporates time-dependent discrimination measures into the feature selection procedure. Specifically, candidate features are evaluated based on their time-dependent AUC across multiple prediction times. This allows the identification of consistent informative variables over time. The proposed approach is implemented within a filter-based selection framework and compared with commonly used embedded methods, including Lasso, Ridge regression, and Elastic Net. The performance of the proposed method is assessed through simulation studies under varying correlation structures and censoring levels, as well as through application to real-world clinical datasets.

Results

Results demonstrate that incorporating time-dependent evaluation into feature selection improves the stability and predictive performance of selected models. In real data, time-dependent selection yields higher predictive accuracy at multiple time points and identifies features with changing importance over time. This work highlights the importance of aligning feature selection criteria with the temporal nature of survival outcomes and provides a practical approach for improving predictive modelling in time-to-event data.

Discussion

Incorporating time-dependent evaluation into feature selection improves alignment between model development and survival outcomes. The proposed framework provides a practical approach for identifying dynamically relevant features and enhances predictive reliability in time-to-event analysis.

[28-P2-17] Assessment of calibration in multistate models for small sample size

Presenter: Lara Rosetta

Authors: Lara Rosetta [1]; Marta Spreafico [1, 2]; Marta Fiocco [1, 2, 3]

Affiliations: [1] Mathematical Institute, Leiden University, Leiden; [2] Leiden University Medical Center, Leiden; [3] Princess Máxima Center for Pediatric Oncology, Utrecht

Introduction

Calibration assessment is well established in statistics, but validated methodologies for multistate risk prediction models remain rather limited.

This study evaluates the performance of methods for assessing the calibration of predicted transition probabilities from multistate models in small sample setting, with a focus on approaches that produce calibration plots.

Methods

We consider methods based on the Aalen-Johansen estimator, pseudo-values, binary logistic recalibration with inverse probability of censoring weights (BLR-IPCW) and multinomial logistic regression with inverse probability of censoring weights (MLR-IPCW).

Their performance in estimating mean and moderate calibration and calibration curves for predicted transition probabilities, was evaluated through a simulation study, focusing on bias and variability.

Data were generated using a clock-forward approach under varying levels of association between the outcome and censoring, assuming conditional independence given predictors.

The methods were further applied to trial data from patients with osteosarcoma to illustrate their practical performance.

Results

Aalen-Johansen, BLR-IPCW and MLR-IPCW methods showed negligible bias when assessing mean calibration the bias remained stable across sample sizes.

In contrast, the variability of calibration plots increased in smaller samples for pseudo-values, BLR-IPCW and MLR-IPCW methods.

In addition, variability was influenced by the number of individuals in each state at the time of evaluation. In particular, states with a low number of individuals and regions corresponding to rare predicted probabilities showed worse performance.

Results under different censoring scenarios were generally consistent across simulations. Increasing the association between the censoring mechanism and the outcome did not substantially affect calibration assessment for any method.

Discussion

These findings provide practical guidance on the application of calibration methods in multistate models, showing the importance of adequate sample size, appropriate choice of evaluation time, and consideration of state occupation. These factors are essential for obtaining reliable calibration estimates in practice.

References: Pate A, Sperrin M, Riley RD, et al. Calibration plots for multistate risk prediction models. Statistics in Medicine. 202443(14):2830-2852. doi: 10.1002/sim.10094

[28-P2-18] Leveraging Relative Survival techniques to enhance the transportability and interpretability of Cox models in metastatic melanoma

Presenter: Toby Hackmann

Authors: Toby Hackmann [1]; Michel Wouters [1, 2, 3]; Ellen Kapiteijn [4]; Ewout Steyerberg [5, 1]; Liesbeth de Wreede [1]

Affiliations: [1] Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden; [2] Department of Surgical Oncology, Netherlands Cancer Institute, Amsterdam; [3] Dutch Institute for Clinical Auditing, Leiden; [4] Department of Medical Oncology, Leiden University Medical Center, Leiden; [5] Julius Center, Utrecht

Introduction

Survival prediction models in oncology often utilize the Cox model to estimate overall survival. However, this approach ignores the distinction between disease-specific mortality and background mortality, which latter part is especially relevant for older patients and long-term outcomes. We aim to demonstrate how a relative survival approach improves treatment effect estimation, model transportability, and the communication of absolute risks.

Methods

We developed a survival prediction model for metastatic cutaneous melanoma, based on observational data from three national registries (the Netherlands (DMTR), Denmark (DAMMED), Slovenia (KrMel) n=6552). We employ an additive relative survival framework where the overall hazard lambdaO(t) is partitioned into the population (background) hazard lambdaP(t), derived from country-specific life tables, and the excess (disease-specific) hazard lambdaE(t x, a) which is estimated for patient i with covariate values x and treatment a using a Cox model:

lambdaO, i(t) = lambdaP, ibigl(t,age,sex,year of start of treatment,countrybigr) + lambdaE, i(t x, a)

We compared this method to a standard Cox model in a leave-one-country-out cross-validation. Additionally, we compared the results of the two methods in simulations where either the relative survival model or the standard Cox model was well-specified.

Results

The additive structure of the model ensured that the treatment effect was applied exclusively to the disease-related excess hazard, rather than to the total hazard. In the prediction model for metastatic melanoma, the absolute value of coefficients for treatment effects (both main effects and interactions) are larger in the relative survival model compared to the standard Cox model. Transportability of the model to new geographic or temporal settings was improved by relaxing the assumption of unchanged all-cause mortality to unchanged disease-related excess mortality.

Discussion

Partitioning hazards via relative survival provides a more precise framework for clinical prediction than standard Cox models. By separating the disease-specific effect from the background mortality, we improve the external validity and transportability of the model. Treatment effects can be estimated on the correct scale of disease-related hazard. Communication of survival predictions is also more nuanced, in particular for older patients, where background mortality can be a large proportion of the total risk.

[28-P2-19] Beyond age and sex: using multi-stakeholder perspectives to define meaningful subgroups for fairness assessment in clinical prediction modelling

Presenter: Molly Wells

Authors: Molly Wells [1]; Jennifer Downing [2]; Mark Goodall [2]; Suzannah Hargreaves [2]; Matthew Sperrin [1]; David Jenkins [1]

Affiliations: [1] University of Manchester, Manchester; [2] University of Liverpool, Liverpool

Introduction

Patient and public involvement and engagement (PPIE) refers to research conducted "with" or "by" members of the public rather than "to", "for", or "about" them. Despite recognition of its importance, PPIE remains underused in the development, validation, and dissemination of clinical prediction models (CPMs). CPMs use patient characteristics (e.g., age, sex) to estimate future outcomes. An emerging area in CPM evaluation is fairness, often defined as equitable predictive performance across patient subgroups. However, guidance is limited on defining clinically meaningful subgroups beyond traditional categories such as age and sex, and on accounting for patient intersectionality. We explored participatory, stakeholder-informed approaches to subgroup identification through patient personas developed with input from patients, carers, and healthcare professionals for a heart failure (HF) CPM.

Methods

Patient personas representing diverse HF patients were initially developed from interviews conducted during model development, with input from patients, carers, and healthcare professionals (e.g., nurses, pharmacists, and GPs). A stakeholder workshop was conducted with participants from

these groups to refine the personas and identify additional subgroups. The aim of these personas was to identify potential subgroups to assess performance and fairness of the HF CPM (RENAL-HF).

Results

The workshop, involving patients, carers, and healthcare professionals, led to adaptation and refinement of existing personas and identification of previously underrepresented subgroups. Participants highlighted gaps in representation of younger HF patients and contributed more detailed, realistic profiles. The personas also revealed clinically relevant intersections of characteristics not typically captured by standard subgroup definitions.

Discussion

There is growing interest in fairness in CPMs, alongside recognition of patient and stakeholder input, but limited guidance on how to utilise these perspectives in practice. This study demonstrates a practical case-study approach using stakeholder-informed personas to define patient subgroups for evaluation of CPM performance. Personas can help identify clinically meaningful and underrepresented subgroups, as well as relevant intersections between characteristics. While some groups may not be well represented in available data (e.g., due to small sample sizes for certain intersections), this approach provides a reproducible framework to support broader inclusion of vulnerable or marginalised groups and enable more robust and sustainable fairness assessment of CPMs.

[28-P2-20] Decision fragility in prediction model development: are decisions robust to uncertainty in predicted risks?

Presenter: Giulia Zamagni

Authors: Giulia Zamagni [1]; Lorenzo Monasta [1]; Giulia Barbati [2]

Affiliations: [1] Institute for Maternal and Child Health - IRCCS 'Burlo Garofolo', Trieste; [2] University of Trieste, Trieste

Introduction

Decision curve analysis (DCA) is increasingly used to evaluate the clinical utility of prediction models. In parallel, uncertainty in individual risk predictions is also recognized as relevant for clinical decision-making. Models with similar discrimination, calibration, and net benefit (NB) may still differ in the stability of individual-level decisions. We therefore conceptualized decision fragility as a complementary measure to NB.

Methods

We developed three models (logistic regression, random forest, and XGBoost) to predict adverse perinatal outcomes in 4,390 women with fetuses small for gestational age. Model performance was assessed by bootstrap optimism-corrected Area Under the Receiver Operating Characteristic Curve (AUC) and Brier score, and by calibration slope calculated in the development dataset. DCA was performed reporting optimism-corrected NB across clinically relevant thresholds. To quantify individual-level uncertainty, bootstrap resampling was used to derive an empirical distribution of

predicted risks for each patient. Uncertainty intervals were defined by the corresponding 2.5th and 97.5th percentiles. At selected thresholds, we defined fragile decisions as those for which the uncertainty interval included the threshold, indicating that small perturbations in the development sample could change the treatment decision. Decision fragility was summarized for each model as the proportion of such cases at the threshold of interest.

Results

The event rate was 8%. Models showed similar performance: optimism-corrected AUC was 0.79 [0.77-0.82] for logistic regression, 0.80 [0.79-0.82] for random forest, and 0.80 [0.78-0.83] for XGBoost. Corrected Brier scores were 0.063 [0.058-0.068], 0.061 [0.056-0.066], and 0.062 [0.057-0.065], respectively. Calibration slopes differed more substantially (1.11, 1.30, and 1.10). NB was comparable across clinically relevant thresholds for example, at 0.08, NB was 0.035 for logistic and random forest, and 0.031 for XGBoost. In contrast, fragility varied: 16%, 21%, and 30% of decisions were unstable, respectively.

Discussion

In model development, optimism-corrected NB summarizes potential clinical utility but does not provide insights into the stability of individual decisions. Decision fragility provides complementary information by quantifying threshold-sensitive predictions and may help distinguish between models with similar NB, particularly when decisions carry important clinical consequences.

[28-P2-21] Quantifying Individual-level Uncertainty in the BOADICEA Breast Cancer Risk Prediction Model

Presenter: Bethan White

Authors: Bethan White [1]; Lorenzo Ficorella [1]; Marc Tischkowitz [1]; Juliet Usher-Smith [1]; Douglas Easton [1]; Antonis Antoniou [1]

Affiliations: [1] University of Cambridge, Cambridge

Introduction

Breast cancer risk prediction models estimate an individual's probability of developing breast cancer and identify those who may benefit most from screening or preventive interventions. The BOADICEA breast cancer risk model, implemented in the CanRisk tool (www.canrisk.org), incorporates genetic, lifestyle, hormonal, family history, and anthropometric data to calculate individualised risk estimates. However, these predictions are subject to model uncertainty, which is rarely quantified at the individual level and can vary considerably depending on risk factor profiles. Quantifying individual-level uncertainty is crucial for evaluating risk prediction models, improving trust in risk estimates, and supporting effective risk communication.

Methods

We developed a framework to quantify individual-level uncertainty in breast cancer risk predictions from the BOADICEA breast cancer risk model. Monte Carlo simulation was used to sample from model parameter uncertainty distributions, and lifetime breast cancer risk was calculated for each sampled parameter set. This produced an uncertainty distribution of the individual's risk, from which summary measures such as uncertainty intervals and the spread across risk categories were calculated.

Results

Uncertainty varied considerably across individuals depending on risk factor values. For a woman with strong family history of breast cancer and median values for other risk factors, the BOADICEA point estimate for lifetime breast cancer risk was 22.3% (moderate risk) with a 95% uncertainty interval of 20.0% to 27.1%. For a woman with strong family history, a high-risk BRCA1 pathogenic variant, and median values for other risk factors, the lifetime risk point estimate was 85.0%, with a wider uncertainty interval of 74.2% to 95.8%.

Discussion

There may be considerable uncertainty in breast cancer risk predictions for some individuals this is particularly relevant when predicted risk lies near clinical decision thresholds. Because uncertainty varies between individuals, modelling it at the individual level is crucial. Communicating uncertainty alongside point estimates can improve transparency and support shared decision-making. This framework also helps identify parameters where improved precision could meaningfully reduce uncertainty in personalised breast cancer risk assessment.

[28-P2-22] Uncertainty quantification for individualized treatment effect prediction in crossover and parallel-group comparison studies

Presenter: Masashi Sawada

Authors: Masashi Sawada [1]; Ryo Emoto [2]; Shigeyuki Matsui [2]

Affiliations: [1] Department of Clinical Biostatistics, Kyoto University School of Public Health, Kyoto; [2] Department of Biostatistics and Data Science, Kyoto University School of Public Health, Kyoto

Introduction

In recent years, many statistical and machine learning methods have been developed for analyzing treatment effect heterogeneity however, although evaluating their quantitative accuracy is essential for assessing the feasibility of precision medicine, such evaluations have not been conducted sufficiently. Previous work demonstrated efficiency gains of self-controlled, crossover (CO) to parallel-group (PG) comparison studies, in estimating treatment effect modifications. Another important direction is uncertainty quantification for individualized treatment effect prediction, particularly in the presence of period effects and carryover effects.

Methods

To quantify predictive uncertainty, we introduced claimability rules on whether an efficacy-related claim at the individual level is statistically supportable based on confidence intervals for individual treatment effects. Covariate profiles that do not satisfy these rules constitute an indeterminate region. We evaluated the impact of sample size, period effects, prognostic heterogeneity, and carryover effects on this region, and identified boundary conditions for CO and PG comparisons became nearly equivalent. We also assessed other supplementary criteria, such as false-claim rate, coverage, and mean squared error across representative scenarios.

Results

Numerical evaluations showed that CO-based prediction tended to yield smaller indeterminate regions than PG-based prediction when prognostic heterogeneity was substantial, reflecting the advantage of within-patient comparison. In contrast, large period effects and carryover effects diminished or, in some scenarios, reversed this advantage. In a practical scenario based on a real-world diabetes study, we identified boundary scenarios yielding indeterminate regions of nearly equivalent size between CO and PG comparisons.

Discussion

These findings suggest that the effectiveness of CO-based individualized treatment effect prediction should be assessed conditionally rather than universally, and that quantitative assessment is necessary to select an analytic strategy for evaluating treatment effect heterogeneity on a case-by-case basis. Another important point is that our study quantitatively demonstrates that a substantial number of patients may fail to meet the criteria for claimability, regardless of whether the comparison is based on CO or PG, thereby highlighting the limitations of predicting treatment outcomes at the individual level.

[28-P2-23] Quantifying the Variability of Explanations: Uncertainty Estimates of SHAP values for Transparent Clinical Machine Learning

Presenter: Simone Buchholz

Authors: Simone Buchholz [1]; Joachim Hegenauer [2]; Ghada El Khawaga [3]

Affiliations: [1] Boehringer Ingelheim Pharma GmbH & Co. KG, Ingelheim am Rhein; [2] Transferzentrum für Digitalisierung, Analytics & Data Science Ulm, Ulm; [3] J. Rettenmaier & Söhne GmbH + Co KG, Rosenberg

Introduction

Artificial Intelligence (AI) increasingly supports drug development and clinical decision making. In parallel, regulators such as EMA and FDA emphasize the need for both quantified predictive uncertainty and transparent model behavior. Ensemble tree-based models satisfy parts of these expectations by offering strong predictive performance and established methods for uncertainty estimation.

SHapley Additive exPlanations (SHAP) has become a widely used approach to interpret these models, providing per instance, feature-level attributions grounded in a solid mathematical framework. However, a critical question remains largely unexamined: how uncertain are the SHAP explanations themselves?

Methods

We extend uncertainty quantification from model predictions to SHAP values. Using the taxonomy of aleatoric and epistemic uncertainty and adapting the methodology of Shaikhina et al. [1], we train ensembles of 1000 models across different model classes and separately estimate aleatoric and epistemic uncertainty. For each model and instance, we compute TreeSHAP explanations, enabling us to quantify SHAP-level uncertainty. Finally, we analyze how attribution uncertainty relates to predictive uncertainty. As a case study, we apply this framework to an open-source heart failure dataset and explore dependencies between prediction-level and SHAP-level variability.

Results

The case study demonstrates that the proposed framework effectively characterizes the uncertainty of SHAP explanations. However, only a small proportion of SHAP variance is explained by variance in predictive uncertainty. Even when models within the ensemble yield highly similar predictions, their SHAP attributions can diverge substantially. This indicates that explanations may be less stable than the predictions they describe.

Discussion

Our findings highlight the need to systematically assess the stability of explanation methods before their use in clinical or regulatory contexts. Understanding the sources of attribution variability is essential for building trust in explanation models intended to guide medical decisions. Further research is required to identify and mitigate factors contributing to SHAP instability.

[1] Shaikhina T, Bhatt U, Zhang R, Georgatzis K, Xiang A, Weller A. Effects of uncertainty on the quality of feature importance explanations. In: AAAI Workshop on Explainable Agency in Artificial Intelligence 2021.

[28-P2-24] A Comparison of Resampling Methods and Their Predictive Performance for Three Machine Learning Algorithms on Imbalanced Medical Data

Presenter: Leandro Teixeira Crespo

Authors: Leandro Teixeira Crespo [1, 2]; Dativa Tibyampansha [2]; Matthias Dreher [3]; Andreas Schwarting [3]; Leszek Wojnowski [4]; Daan Apeldoorn [2, 5]; Andre Thevapalan [2]; Felix Schuhknecht [1]; Gabriele Kern-Isberner [6]

Affiliations: [1] Institute of Computer Science, Johannes Gutenberg University, Germany, Mainz; [2] Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI), University Medical Center of the Johannes Gutenberg University Mainz, Mainz; [3] Division of Rheumatology and Clinical Immunology, Department of Internal Medicine I, University Medical Center of the Johannes Gutenberg University, Mainz, Germany, Mainz; [4] Institute of Pharmacology, Mainz; [5] IU International University of Applied Sciences, Düsseldorf; [6] TU Dortmund University, Dortmund

Introduction

In this work, we analyze the effectiveness of different resampling methods and combinations thereof on imbalanced medical data for three ML algorithms, based on [3].

Methods

We investigated four Resampling methods:

Synthetic Minority Over-sampling (SMOTE [2]) Random Oversampling (ROS) Tomek Links (TOMEK [4]) Edited Nearest Neighbours (ENN [5]). In ROS, existing samples are randomly duplicated, while SMOTE creates synthetic samples between randomly selected samples in the dataset and their nearest neighbors. TOMEK removes similar samples with different labels, while ENN removes samples that differ in their label compared to the majority of similar samples. We used an imbalanced medical dataset obtained from a survey of 815 patients screened for three rheumatic diseases (Psoriatic arthritis (PsA), Rheumatoid arthritis (RA) and axial spondyloarthritis (SpA) 25% diseased, 75% not diseased). We trained three different models: a neural network (NN), a decision tree (DT), and a hierarchical knowledge base (HKB knowledge represented through rules [1]).

Results

Figure 1: One-vs.-Rest Sensitivity and Specificity for tested resampling methods and models.

The results (Figure 1) show that resampling increases the sensitivity for the minority classes (diseased), with a decrease in sensitivity for the majority class (No Disease) and accuracy. SMOTEENN performed best for NN and DT with an average minority sensitivity of 0.47 and 0.39 respectively, SMOTE performed best for HKB with an average minority sensitivity of 0.58.

Discussion

Through Resampling, improved minority sensitivity was achieved at the cost of accuracy and majority sensitivity, with marginal impact on minority specificity. This represents a strategic prioritization of disease detection by the Resampling techniques. These results are potentially limited by the small sample size.

[1] Apeldoorn D, Panholzer Torsten. Automated Creation of Expert Systems with the InteKRator Toolbox. Stud Health Technol Inform. 2021283:46-55. DOI: https://doi.org/10.3233/SHTI210540

[2] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. JAIR. 200216:321-357. DOI: https://doi.org/10.1613/jair.953

[3] Teixeira Crespo L. Vergleich unterschiedlicher Vorhersageverfahren auf Rheumadaten. Masterarbeit. TU Dortmund University. 2025.

[4] Tomek I. Two Modifications of CNN. IEEE Trans Syst Man Cybern. 1976SMC-6(11):769-772. DOI: https://doi.org/10.1109/TSMC.1976.4309452

[5] Wilson DL. Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. IEEE Trans Syst Man Cybern. 1972SMC-2(3):408-421. DOI: https://doi.org/10.1109/TSMC.1972.4309137

P3 - Clinical Trials

Poster area: P3

[28-P3-01] Design-stage recruitment analysis requirements for clinical trial protocols submitted to public funding organizations: A mixed-methods scoping review

Presenter: Philip Heesen

Authors: Philip Heesen [1]; Katarina Zatkova [1]; Malgorzata Roos [1]

Affiliations: [1] University of Zurich, Zurich

Introduction

Adequate participant recruitment is essential for clinical trial success, yet many publicly funded trials fall short of accrual targets, leading to delays, underpowered results, or premature termination. Design-stage recruitment analysis (DSRA) provides the methodological foundation for estimating feasible recruitment rates and timelines before a trial begins. Despite their importance, DSRA requirements across public funding agencies are inconsistent and lack systematic comparisons. This scoping review maps and synthesizes DSRA-related requirements across major public funders to identify areas of convergence and divergence and to inform a unified DSRA reporting framework.

Methods

A mixed-methods scoping review was conducted following a pre-registered protocol. Funding agencies from 194 countries were screened through systematic Google searches and direct screening of international funder directories. Eligible instructions for clinical trial protocol were publicly accessible and written in English. These instructions had to minimally address sample size and recruitment duration. Two reviewers independently extracted DSRA-related requirements into a standardized form. Extracted items were categorized into medical, statistical, and administrative components and quantified by the proportion of agencies requiring each item.

Results

Of the 194 countries screened, four agencies met all inclusion criteria: Swiss National Science Foundation (SNF), Deutsche Forschungsgemeinschaft (DFG), Austrian Science Fund (FWF), and National Institutes of Health (NIH). Sample size was the only requirement mentioned by all four agencies (100%). Trial duration was required by 75%, and recruitment duration by 50-75%. Recruitment

feasibility explanation, recruitment strategy, data sources for assessing recruitment potential, and schematic recruitment diagrams were each mentioned by 50%. Items such as pilot data, anticipated accrual rates, and patient availability were required by only one agency (25%). Statistical requirements had greater cross-agency coverage than medical and administrative requirements.

Discussion

DSRA requirements vary substantially across public funders, with only sample size universally required. These findings highlight the need for a harmonized DSRA reporting template integrating statistical, medical, and administrative components. Such standardization could improve trial feasibility, reduce research waste, and support investigators in meeting cross-agency expectations. An accompanying open-source software implementation is being developed to facilitate reproducible DSRA in practice.

[28-P3-02] Scaling recruitment rates made easy: Open-source Shiny application for recruitment time and participant count predictions using Erlang and Poisson models

Presenter: Alvin Walker

Authors: Alvin Walker [1]; Katarina Zatkova [1]; Manuela Ott [1]; Malgorzata Roos [1]

Affiliations: [1] University of Zurich, Zurich

Introduction

Inaccurate recruitment analyses can lead to inappropriately designed studies that fail to recruit enough participants in a timely manner, resulting in inconclusive statistical findings and study discontinuations. Predictions provided by probabilistic models, such as the Erlang model for recruitment time and the Poisson model for participant counts, reduce the risk of inaccurate analysis by enabling decisions that account for uncertainties. Despite their value, the statistical complexity of these methods creates implementation challenges for medical study designers, principal investigators, and study managers. Already during the planning phase, different administrative scenarios for increasing recruitment rates should be considered to reduce the allotted time for recruitment. Recruitment rates can then also be scaled by the retention rate to predict the sample size available upon study completion. Furthermore, in ongoing studies with lower recruitment rates than anticipated, additional administrative actions should be considered to increase those rates and help meet the study timeline. These operations involve scaling of recruitment rates. However, transforming probabilistic predictions to account for this scaling remains complex and requires specialized knowledge or software.

Methods

The open-source R programming language was used to develop a reliable and user-friendly Shiny application for recruitment analysis based on scaled recruitment rates.

Results

The open-source Shiny application enables users to input and dynamically scale recruitment rates in Erlang and Poisson models and provides predictive distributions, estimates, figures, and interpretations. Users can export the results to a PDF and submit them to a funding agency with a study protocol.

Discussion

This Shiny application offers study designers and managers, principal investigators, and funding agencies direct access to probabilistic predictions of time and counts, eliminating the need for independent scaling implementations. By showing the impact of hypothetical scaling of recruitment rates on time and count predictions, the Shiny application helps ensure that medical studies are planned realistically. It thus helps increase the number of successfully completed studies that provide statistically reliable findings based on sufficient participant counts.

[28-P3-03] Anticlustering as a new technique for allocating patients into treatment groups in clinical trials

Presenter: Carolin Herrmann

Authors: Carolin Herrmann [1]; Martin Papenberg [2]

Affiliations: [1] Mathematical Institute and Center for Digital Medicine, Heinrich Heine University Düsseldorf, Düsseldorf; [2] Institute for Experimental Psychology and Center for Digital Medicine, Heinrich Heine University Düsseldorf, Düsseldorf

Introduction

Comparability between treatment groups is a central quality criterion of clinical trials. Hence, randomization, minimization, and variations thereof, have been applied for decades. The different approaches come with advantages and shortcomings, often influenced by aspects such as sample size, recruitment pattern, number of centers and other known important prognostic variables.

Methods

We introduce anticlustering as a new technique for allocating patients to treatment groups in clinical trials. The algorithm stems from the field of bioinformatics (originally going back to [1]) and is publicly available in the R package anticlust. To apply anticlustering to clinical trials, we modified the algorithm slightly: We extended the method both with respect to making patient inclusions at different time points possible and allocating patients in blocks. Afterwards, we conducted a comprehensive simulation study to evaluate the balance of prognostic variables between two treatment groups. We considered both categorical and continuous prognostic variables and compared anticlustering with simple randomization, block randomization, stratified block randomization, minimization and dynamic block randomization.

Results

Anticlustering works similarly to dynamic block randomization or minimization when considering the quantification of group balances. However, it uses a different algorithm for maximizing the balance, almost always leading to an outperformance of the other approaches. We evaluated the discrepancy between treatment groups (%) for categorical variables and the mean absolute difference between group means for continuous variables depending on the number of covariates. For both settings, simulation results were very similar with anticlustering outperforming the other approaches listed in the methods section.

Discussion

Anticlustering is a promising new approach for allocating patients to treatment groups in clinical trials. Due to its construction, it provides great flexibility with respect to different application scenarios and outperforms many standard techniques. At the same time, anticlustering is computationally fast compared to the other approaches considered here and is available in the R package anticlust.

Literature:

[1] Späth, H. (1986). Anticlustering: Maximizing the variance criterion. Control and Cybernetics, 15(2), 213-218.

[28-P3-04] Statistical Methods and Regulatory Considerations in Phase I Clinical Trials

Presenter: Tzu-Chuan Lin

Authors: Tzu-Chuan Lin [1]

Affiliations: [1] Division of New Drugs, Center for Drug Evaluation, Taipei City

Introduction

This presentation provides an overview of Phase I clinical trial design with a focus on dose-escalation methodologies. Key design elements include study population selection, justification of the starting dose, prespecified definitions of dose-limiting toxicities, and structured dose-decision frameworks across rule-based, model-based, and model-assisted approaches.

Methods

A simulation comparison of 3+3, CRM, and BOIN across multiple toxicity scenarios.

Results

A simulation comparison of 3+3, CRM, and BOIN across multiple toxicity scenarios shows that model-based and model-assisted designs achieve higher accuracy and efficiency, whereas the 3+3 design remains conservative with limited information gain.

Discussion

From a regulatory perspective, priority is placed on early detection of DLTs, availability of overdose-control mechanisms, and transparent operating characteristics across scenarios. Additional expectations include predefined escalation-hold triggers, cohort-level safety oversight, and ensuring that RP2D determination integrates MTD with PK/PD exposure-response data, biomarkers, and mechanistic rationale rather than relying solely on MTD.

[28-P3-05] Comparison of two modelling approaches for two-population phase I/II dose-finding trials

Presenter: Jan Meyring

Authors: Jan Meyring [1, 2]; Lucie Biard [1, 2]; Moreno Ursino [3]; Sylvie Chevret [1, 2]

Affiliations: [1] ECSTRRA Team, IRSL, INSERM UMR1342, Université Paris Cité, Paris; [2] Institut de Recherche Saint-Louis, Université Paris Cité, Paris; [3] HeKA (UMR 1346) INRIA/INSERM/Université Paris Cité, Paris

Introduction

Phase I trials identify the maximum tolerated dose, assuming toxicity and efficacy increase with dose. For new anti-cancer agents, efficacy may plateau and few toxicities occur at higher doses. Phase I/II designs address this setting. Most phase I/II designs assume population homogeneity and a common optimal dose, though subpopulations may exhibit differential treatment response. We propose two designs for identifying subgroup-specific optimal doses with shared information: a pairwise borrowing approach and Bayesian hierarchical modelling. We focus on only two prespecified subgroups which correspond to a common and practical situation in early-phase clinical trials with limited sample sizes. The objective is to identify the optimal biological dose (OBD), defined as the highest efficacious dose that does not exceed a prespecified toxicity threshold.

Methods

Both methods were derived from a phase I/II model-based design with binary outcomes for homogeneous populations, using features of the continual reassessment method and Bayesian model averaging to account for various efficacy relationship shapes. For pairwise borrowing, the toxicity and efficacy parameters are first estimated separately for each subgroup and outcome. Then, subgroup similarity is assessed using a probability divergence measure to determine whether information should be shared between the two subgroups this allows borrowing for either outcome, both, or neither. If borrowing occurs, the estimation is rerun assuming a homogeneous population. In the Bayesian hierarchical model, a hyperprior on the dose-toxicity and dose-efficacy models allows information to be shared across subgroups. We evaluated the operating characteristics of the designs in various scenarios through simulations, comparing them to benchmark approaches (two separated models and pooled analysis).

Results

While the proposed Bayesian hierarchical model did not outperform the separated models, the pairwise borrowing approach demonstrated favourable performance in common scenarios while struggling with more difficult ones. It effectively shared information when subgroups had similar toxicity profiles but struggled to share information on efficacy, due to the higher variability in responses.

Discussion

Although the proposed pairwise borrowing design provided a parsimonious approach to subgroup-specific dose finding in phase I/II trials with heterogeneous populations, its performance may be limited in settings where efficacy-based information sharing is required.

[28-P3-06] Level Set Estimation for Identifying Acceptable Doses in Phase I/II Trials

Presenter: Hiroki Katsumata

Authors: Hiroki Katsumata [1]; Kota Matsui [2]; Shigeyuki Matsui [2]

Affiliations: [1] Depatment of Biostatistics, Nagoya University School of Medicine, Nagoya City; [2] Department of Biostatistics and Data Science, Kyoto University School of Public Health, Kyoto

Introduction

Modern targeted and immuno-oncology agents may exhibit meaningful activity without a monotone "more dose, more toxicity and more efficacy" relationship. Hence, selecting a single recommended dose solely from toxicity (e.g., a maximum tolerated dose) can be suboptimal. Many existing approaches combine toxicity and efficacy into a one-dimensional utility function, but this requires subjective weighting and may obscure the trade-off between the two outcomes. We proposed a method of level set estimation (LSE) that aims to identify the entire set of doses that simultaneously satisfy prespecified tolerability and efficacy requirements while retaining the two-dimensional structure of the outcomes. By estimating this acceptable-dose region, clinicians can select one or multiple recommended doses based on overall benefit-risk considerations.

Methods

Candidate doses are scaled on [0,1]. We define a target dose set (TDS) as those with toxicity and efficacy probabilities respectively below and above predefined thresholds. Both dose-response curves are modeled using independent Gaussian process priors, with posteriors updated via Laplace approximation. To determine the next cohort dose, we maximize an acquisition score prioritizing doses with the highest classification uncertainty. These uncertainties for toxicity and efficacy are combined multiplicatively to encourage exploration of the two-dimensional toxicity-efficacy space. Evaluation utilizes a set-estimation loss penalizing discrepancies between estimated and true TDS, weighted by distance from thresholds.

Results

In simulations (toxicity threshold 0.30 efficacy threshold 0.50 cohort size 3 12 cohorts/36 patients 100 replicates 12 scenarios), the LSE-based design was compared with Bayesian optimization, STEIN, and BOIN-ET. When multiple acceptable doses existed, the proposed method allocated patients more evenly, reflecting its goal of estimating a dose set rather than a single optimum. According to the proposed loss metric, the LSE method performed comparably to or better than existing approaches.

Discussion

The results support LSE as a principled framework for learning an acceptable-dose region under joint toxicity-efficacy constraints without collapsing outcomes into a single utility measure. However, the method sometimes selected doses above the true target set, motivating additional safety-oriented constraints. Further investigation of acquisition strategies and hyperparameter choices is warranted.

[28-P3-07] A Safe Bayesian Optimization Design for Phase I/II Dose-Finding Trials in Oncology

Presenter: Kenta Wakabayashi

Authors: Kenta Wakabayashi [1]; Kota Matsui [2]; Shigeyuki Matsui [2]

Affiliations: [1] Department of Clinical Biostatistics, Kyoto University School of Public Health, Kyoto; [2] Department of Biostatistics and Data Science, Kyoto University School of Public Health, Kyoto

Introduction

The primary objective of phase I/II clinical trials for modern oncology drugs, such as immune checkpoint inhibitors, is to identify an appropriate dose by integrating toxicity and preliminary efficacy outcomes. Many recently proposed designs address the efficacy-toxicity trade-offs by employing a one-dimensional utility function that integrates these factors based on subjective weightings however, reliance on utility functions raises concerns regarding interpretability, the arbitrariness of recommended doses, and the difficulty in reaching consensus.

Methods

We develop a new design method based on the framework of safe Bayesian optimization, which identifies the optimal dose (OD) that maximizes the probability of efficacy while keeping the probability of toxicity below a certain threshold. Specifically, this method requires an initial safe set, comprising pre-specified safe doses, and administration is initiated within this range. In subsequent steps, the safe set is updated by augmenting the initial safe set with doses highly likely to satisfy the toxicity tolerance level. Restricting dose selection to this set ensures participant safety.

Results

Under existing methods, simulations revealed that the weight parameters of the utility function tended to moderately influence the selection frequency of the true OD. Across multiple scenarios involving various toxicity and efficacy profiles, the proposed method achieved true OD selection

frequency comparable to existing methods in several of these scenarios, it also reduced the number of participants assigned to excessively toxic doses. However, when the true OD lay within the high-dose range, it tended to recommend overly conservative doses, resulting in lower estimation accuracy.

Discussion

Compared to existing methods integrating toxicity and efficacy into a one-dimensional utility function, the proposed approach reduces the arbitrariness of the recommended doses and improves clinical interpretability. Furthermore, it is expected to effectively identify safe and efficacious doses in multiple phase I/II oncology trials. Despite these advantages, a key limitation of the proposed method is that misspecification of the initial safe set increases the risk of exposing participants to overly toxic doses because doses in this set are subject to exploration throughout the trial. Although such risks can be mitigated through appropriate safety monitoring frameworks, prior validation of this set is essential for clinical implementation.

[28-P3-08] Bayesian Seamless Phase I/II Dose-Finding and Probability of Success-Based Asset Selection in Early Oncology Development

Presenter: Dimitris Kontos

Authors: Dimitris Kontos [1]; Ming-Dauh Wang [2]; Alberto Pessia [3]; Robert Adams [4]

Affiliations: [1] BCC AG, Pharmaceuticals, Basel; [2] Bayer Healthcare Pharmaceuticals Inc., East Hanover; [3] Bayer Oy, Espoo; [4] Bayer Aktiengesellschaft, Berlin

Introduction

In traditional oncology drug development, safety and therapeutic activity are evaluated separately, with limited information sharing between Phase I dose-finding and Phase II efficacy evaluation. A seamless Phase I/II design simultaneously targets the Optimum Biological Dose (OBD), maximizing information sharing across safety and efficacy endpoints. In multi-cycle oncology treatments, both outcomes may have delayed onset. We present an integrated Quantitative Decision Making (QDM) framework applied to two first-in-human studies in the same oncology indication, both employing radio-labeled Targeted Alpha Therapy (TAT) directed at the same validated molecular target: one utilizing a targeted monoclonal antibody, the other a targeted small molecule.

Methods

We applied the Joint Time-to-Event Continual Reassessment Method (Joint TITE-CRM), a model-based Phase I/II design accommodating delayed onset of binary safety and efficacy outcomes across multiple treatment cycles. Probability of Success (PoS) estimates were computed as posterior predictive probabilities that the Recommended Phase II Dose (RP2D) would satisfy predefined joint Go/No-Go criteria on tolerability and efficacy at the next development stage, enabling probabilistic asset selection and competitor benchmarking across both compounds.

Results

The Joint TITE-CRM identified distinct OBD recommendations for the monoclonal antibody and the small molecule, with delayed outcome weighting materially influencing dose escalation decisions. PoS estimates derived from posterior distributions enabled transparent probabilistic differentiation between the two compounds against Go/No-Go criteria, supporting a data-driven asset selection recommendation. Benchmarking of both RP2D candidates against external competitor data via PoS provided additional strategic context for the Go/No-Go decision. Graphical outputs facilitated real-time communication of model recommendations to cross-functional clinical teams throughout both studies.

Discussion

This integrated QDM framework - combining Joint TITE-CRM for seamless Phase I/II dose-finding with Bayesian PoS for asset selection and competitor benchmarking - demonstrates the value of evidence-driven statistical decision making in early TAT oncology development. Limitations include binary outcome definitions and monotone dose-response assumptions. The framework offers a generalizable template for early oncology programs with delayed outcomes and multi-asset selection needs.

References: [1] Barnett H, Boix O, Kontos D & Jaki T (2025). Joint TITE-CRM: A Design for Dose Finding Studies for Therapies with Late-Onset Safety and Activity Outcomes. Stat Biopharm Res, 17(1):149-160. https://doi.org/10.1080/19466315.2024.2333388

[28-P3-09] A Non-parametric Approach to Simultaneous Efficacy Detection and Estimation of the Dose-Response Relationship in Phase II Trials

Presenter: Riko Asano

Authors: Riko Asano [1]; Kota Matsui [2]; Shigeyuki Matsui [2]

Affiliations: [1] Department of Clinical Biostatistics, Kyoto University School of Public Health, Kyoto; [2] Department of Biostatistics and Data Science, Kyoto University School of Public Health, Kyoto

Introduction

In phase II clinical trials, the detection of drug efficacy signals and the estimation of dose-response relationships are critical processes for decision-making in clinical development. Although approaches that combine multiple testing procedures with parametric dose-response modelling, such as MCP-Mod, are frequently employed, their performance relies heavily on a pre-specified set of candidate models, posing risks of bias and reduced accuracy under model misspecification. Furthermore, such approaches lack flexibility when unexpected response patterns are observed. To address these issues, we propose a unified, data-driven analysis framework based on Gaussian processes (GP) that integrates estimation and detection without being restricted by specific functional forms.

Methods

The proposed method consists of two phases: estimation and detection. In the estimation phase, we consider two candidate models: a standard GP and a shape-constrained model (monotonic spline plus a GP term). We employ a data-splitting approach, where one portion of the data is used to adaptively select the better-fitting model, and the remaining portion is used for final curve estimation. In the subsequent detection phase, a randomization test is performed based on the estimates from the selected model to assess the presence of a dose-response signal. We conduct simulation studies to compare the performance of the proposed and MCP-Mod approaches under various dose-response shapes.

Results

Simulation results indicated that under ideal conditions, where the true dose-response shape was included in the candidate model set, MCP-Mod exhibited high efficiency by leveraging its parametric properties. Conversely, in scenarios involving model misspecification, the proposed method significantly reduced the mean square error in estimating the dose-response curve. Furthermore, the randomization test maintained high statistical power even when parametric assumptions were violated.

Discussion

Since the underlying drug response profiles remain largely uncharacterized in many phase II trials, the inherent flexibility of the proposed non-parametric approach, which does not rely on predefined functional forms, is likely to reduce potential estimation bias and provide a safeguard against model misspecification. Consequently, this approach is expected to serve as an alternative comprehensive analysis package for providing robust and reliable dose-response information for decision-making in clinical development.

[28-P3-10] Statistical methods to reduce Selection Bias in Dose-Finding Studies with Binary Endpoints

Presenter: Alexandra Balzer

Authors: Alexandra Balzer [1]; Meinhard Kieser [2]

Affiliations: [1] Institute of Medical Biometry, University Heidelberg, Heidelberg; [2] Institute of Medical Biometry, University of Heidelberg, Heidelberg

Introduction

In oncology drug development, phase II dose-finding studies are essential to identify promising dose levels for confirmatory phase III trials. To address the challenge of dose optimization, the FDA's Oncology Center of Excellence launched the OPTIMUS project. A key challenge is the unbiased estimation of therapeutic effects, as true treatment effects are often overestimated in phase II and cannot be confirmed in phase III trials. While bias-correction methods for dose-finding trials have been presented for normally distributed outcomes [1], none have been presented for binary endpoints in

this context, which are common in oncology (e.g., tumor response). This work evaluates bootstrap-based and Bayesian hierarchical methods for reducing selection bias in dose-finding studies with binary endpoints.

Methods

A simulation study was conducted considering nine different underlying dose-response relationships with five dose levels and 40 patients per dose.

First, single and double (non-)parametric bootstrap methods were evaluated to compare different numbers of bootstrap repetitions based on bias and mean squared error (MSE) for the true maximal dose, across the methods and dose-response relationships.

For Bayesian hierarchical models, various hyperpriors for the variance parameter were investigated, including Gamma, Half-t, and Uniform distributions with different parameterizations.

Results

Regarding the bootstrap-based methods, the number of bootstrap repetitions had no effect on bias or MSE across all dose-response scenarios. While the single bootstrap yielded a positive bias on average, the double bootstrap method tended to have a negative bias. The MSE was higher for double bootstrap compared to the single bootstrap. For the Bayesian hierarchical models, the Gamma and Uniform hyperpriors under suitable parameterizations achieved the best results in terms of bias and MSE on average across all dose-response scenarios. The Gamma and Half-t hyperpriors were sensitive to their parameterization, whereas the Uniform hyperprior demonstrated greater robustness.

Discussion

Bayesian model selection should consider hyperprior choice and parameterization carefully, while bootstrap methods perform well with few repetitions. Future work will examine additional scenarios, including fewer dose levels and smaller sample sizes per dose.

[1] Zhan T. A class of computational methods to reduce selection bias when designing Phase 3 clinical trials. Statistics in Medicine. 202443(10):1993-2006. doi: 10.1002/sim.10041

[28-P3-11] Assessing the Efficacy-Safety Conflict at Early Efficacy Stopping in Group Sequential Designs: A Simulation-Based Evaluation

Presenter: Mohammad Vossoughi

Authors: Mohammad Vossoughi [1]; Alexander Pieper [1]

Affiliations: [1] Evidenze Germany GmbH, Essen

Introduction

Group Sequential Designs (GSDs) allow clinical trials to stop early before the planned end of the study. However, if a trial stops early due to efficacy, fewer safety events may have been observed than originally planned. As a result, the absence of observed safety issues at the time of stopping does not necessarily mean that the treatment is proven to be safe.

Methods

We defined a safety threshold rule, which would prevent a study from being stopped for efficacy, if the upper bound of the confidence interval for the expected safety event exceeds the predefined safety threshold. To investigate this rule, we developed a Shiny app allowing study simulations for various types of GSDs, timings and number of interim analyses (IAs) and the expected rate of safety events. For each scenario, we estimated how often trials that stopped early for efficacy accumulated enough safety information to rule out unacceptable safety risks.

Results

Overall, adding a safety requirement alongside the efficacy criterion has a small impact on the study's preplanned power. This effect on loss in power depends on the strictness of the safety threshold. For many study scenarios we see almost no chance of early stopping because of overwhelming efficacy-safety conflicts at early stages. For most cases it is recommendable to reduce the number of IAs to one IA at a late-stage information. In the case of low-tolerance safety thresholds trials, where we would like to see strictly the expected rate of safety event, Pocock's advantage of having many stops in the early stages is lost due to many efficacy-safety conflicts. For these scenarios the O'Brien-Fleming design or a single late-stage IA is preferable to ensure that the safety data is sufficiently mature and precise to meet regulatory demands.

Discussion

Introducing clear safety guidance for interim decisions may reduce the likelihood of efficacy-based stopping of trials comprising safety uncertainty. However, it provides an important advantage: decision rules can be defined before the study begins, considering a meaningful number and timing of IAs and reducing the need for ad-hoc discussions. This can support more transparent and consistent decision-making in clinical trials.

[28-P3-12] Current practices in the design of master protocol trials: a scoping review

Presenter: Melissa Middleton

Authors: Melissa Middleton [1]; Chris Selman [2]; Alistair McLean [3]; Leon Di Stefano [1]; Julie Simpson [3]; Robert Mahar [3]; Katherine Lee [1]

Affiliations: [1] Murdoch Children's Research Institute, Melbourne; [2] Murdoch Children's Research Institute, Parkville; [3] University of Melbourne, Melbourne

Introduction

Master protocol trial designs enable simultaneous evaluation of multiple interventions and/or intervention effects across subpopulations within a single overarching protocol. Although their use is increasing, it is unclear whether these trials are adhering to best statistical practice given their complex nature. This scoping review aimed to describe current practice in the statistical design and analysis planning of master protocol trials and identify areas for improvement

Methods

We conducted a scoping review of master protocol trials that registered one or more components (e.g. master protocol, intervention, subpopulation) on ClinicalTrials.gov between January 2022 and December 2024. Statistical documentation (e.g., protocols, statistical analysis plans) was sourced from ClinicalTrials.gov, PubMed, trial websites and the arXiv preprint repository. We extracted information on statistical design (e.g. use of estimands, sample size calculations) and the planned analysis (e.g. handling of non-concurrent controls and missing data).

Results

We identified and screened 498 registered trial components resulting in 117 eligible trials for inclusion and data extraction. Platform trials were the most prevalent self-identified master protocol design (51%). The availability of trial documentation was limited with only 31 (26%) trials having protocol documents publicly available. Of the 31 trials with available documentation, the estimands framework was implemented in 8 trials, and the planned handling of intercurrent events was inconsistently described, with several trials outlining expected intercurrent events but specifying analysis approaches that implicitly assumed their absence. How trials planned to handle missing data was not provided for 13 trials.

Discussion

This review highlights the gaps in transparent statistical planning in master protocol trials. Publicly available documentation was limited, and when accessible, often lacked sufficient detail to assure trials were following current recommendations in terms of statistical rigour. These findings likely reflect both the inherent complexity of master protocol designs and the absence of tailored, widely adopted guidance for their statistical documentation. There is a need for guidance on how master protocol trials should adhere to current best statistical practice, how statistical information in master protocols should be structured, and which documents should be made publicly available, and at what stage of the trial.

[28-P3-13] How to optimize dynamic borrowing in basket trials - Results of a comparison study

Presenter: Lukas D Sauer

Authors: Lukas D Sauer [1]; Alexander Ritz [2]; Meinhard Kieser [1]

Affiliations: [1] Institute of Medical Biometry, University of Heidelberg, Heidelberg; [2] Institute of Mathematics, TU Clausthal, Clausthal-Zellerfeld

Introduction

In oncological drug development, clinical trials often face a dilemma between two requirements. Sample sizes tend to be small, e.g. for drugs targeted at specific genetic traits. At the same time, one may test the drug in different cancer localizations (e.g. lung, breast, and colon cancer) if its treatment mechanism is tissue-agnostic. This two-fold challenge can be addressed by investigating the different indications in one single trial. Such a trial in which one treatment is tested in a variety

of subgroups is called a basket trial. Statistically, this subgroup structure can be leveraged by using borrowing: The basic idea is that subgroups responding similarly to the treatment will "borrow" data from one another, whereas subgroups responding differently will be analyzed separately. Many borrowing techniques have been suggested in the literature. However, as usual, there is no uniformly favorable approach: Power gains in subgroups with good response to treatment coincide with type-I error inflation in subgroups not responding. An objective investigation of optimizing the trade-off between benefits and disadvantages of borrowing mechanisms in basket trials is still missing.

Methods

Optimizing borrowing in a basket trial encompasses the following steps: defining a set of outcome scenarios for the subgroups, defining a target for optimization in form of a utility function, and numerical optimization of said function. We conducted a simulation study in order to investigate the effect of different utility functions in optimizing an exemplary Bayesian basket trial design suggested by Fujikawa et al. 2020. The utility functions we consider are the expected number of correct decisions used by Broglio et al. 2022, a power type-I error difference used by Jiang et al. 2021, and other candidates.

Results

Both utility functions suggested by Jiang et al. and by Broglio et al. showed good performance for tuning borrowing in a range of different scenarios. We discuss their pros and cons in detail.

Discussion

Implementing information borrowing in a basket trial is challenging, because one deliberately introduces bias hoping for power gains in some scenarios. Our optimization framework for basket trials aids in tuning this trade-off, thus helping in informed trial planning.

[28-P3-14] Treatment effect evaluation with acceptable non-concurrent controls in platform trials

Presenter: Siyuan Ren

Authors: Siyuan Ren [1]; Takashi Omori [2]; Isao Yokota [1]

Affiliations: [1] Department of Biostatistics, Faculty of Medicine, Hokkaido University, Sapporo; [2] Department of Clinical Biostatistics, Graduate School of Medicine, Kyoto University, Kyoto

Introduction

Platform trials enable efficient evaluation of multiple investigational treatments within a single study by sharing a common control group. When evaluating a newly added treatment arm, incorporating non-concurrent control data can increase statistical power. However, temporal changes in patient populations or trial environments may introduce heterogeneity between non-concurrent and concurrent controls, leading to biased treatment effect estimates. Several analytical methods that fully borrow non-concurrent control data have been proposed to address this issue, but they may not sufficiently account for increased heterogeneity when earlier control data are temporally distant from the concurrent period.

Methods

We developed a method to evaluate treatment effects in platform trials by selectively incorporating non-concurrent control data regarded as comparable to concurrent controls. The proposed method divides non-concurrent control data into time intervals and sequentially identifies acceptable intervals using likelihood ratios, starting from the interval closest to the concurrent control period and moving backward. Treatment effects are evaluated using posterior distributions of cure proportions based on the selected control data and treatment group data.

Results

The six evaluated methods were classified into two types: those yielding unbiased treatment effect estimates with non-inflated alpha error but power similar to analyses using only concurrent control data, and those yielding higher power at the cost of biased estimates and inflated alpha error. The proposed method belonged to the latter group but demonstrated smaller bias and less alpha error inflation than the other methods in this category.

Discussion

The proposed method can evaluate treatment effects under both the presence and absence of true effects and temporal heterogeneity, although it may slightly inflate the alpha error.

[28-P3-15] Counting What Counts: Novel Adaptive Multi-Arm Multi-Stage (MAMS) designs for Poisson and Negative Binomial Count Outcomes

Presenter: Cornelia Horsch

Authors: Cornelia Horsch [1]; Yumeng Liu [2]; Mahesh Kb Parmar [3, 4]; Benjamin Mayer [1]; Babak Choodari-Oskooei [5]

Affiliations: [1] Universität Ulm, Institut für Epidemiologie und Med. Biometrie, Ulm; [2] University of Bristol, Bristol; [3] UCL Innovative Clinical Trial Unit, London; [4] Institute of Clinical Trials and Methodology at UCL, London; [5] UCL, London

Introduction

The multi-arm, multi-stage (MAMS) trial design is an efficient adaptive approach which tests multiple experimental treatments in parallel and compares them separately against a common control arm in several stages, ceasing recruitment early to those treatments which appear futile.

The existing methods (and software) allow for continuous, binary and time-to-event clinical outcomes. However, these do not allow for (overdispersed) count outcomes such as the number of infections after vaccination. We propose a novel MAMS design to allow for these more complex outcome types. We implemented the new design in user-friendly software to facilitate the design of such complex studies.

Methods

We present the design, which allows for the use of an information-rich intermediate outcome at interim stages, and outline the steps to design a MAMS trial with count outcomes. The new design is then applied to the example 3-arm COLO-PREVENT trial, investigating interventions for colorectal polyp prevention. Finally, we consider additional design features which could make the approach even more efficient, such as using an intermediate outcome to weed out potentially ineffective interventions at earlier stages.

Results

The primary outcome in the COLO-PREVENT trial follows an overdispersed negative binomial distribution. The summary outcome measure is the mean number of polyps per patient within one year. We illustrate how such a trial can be designed more efficiently using this new adaptive approach, targeting pre-specified overall familywise (and pairwise) Type I error rates and power. We present the projected trial timelines and required sample sizes, and compare them to alternative designs. The new designs are adaptive and perform better than the alternative designs, with much smaller expected sample sizes.

Discussion

This paper introduces a novel MAMS framework to facilitate the design of trials with a (potentially overdispersed) count outcome. The new MAMS design reduces the average sample size by 41% compared with traditional designs. Furthermore, it offers higher probability of identifying better treatments reduction in average cost and overall trial duration by testing multiple treatments within one framework and a speedy evaluation of new interventions.

[28-P3-16] From Hybrid to Fully Bayesian: A Simulation Study Evaluating Operating Characteristics and Efficiency in an Adaptive Platform Trial

Presenter: Jose Manuel Lema Martinez

Authors: Jose Manuel Lema Martinez [1, 2]; Kevin Wilson [2]; Nina Wilson [3]

Affiliations: [1] University of Oxford, Oxford; [2] Newcastle University, newcastle upon tyne; [3] Population Health Sciences Institute, Newcastle University, newcastle upon tyne

Introduction

The adoption of fully Bayesian adaptive trial designs has been limited by regulatory concerns over frequentist operating characteristics, particularly type I error control. This has driven the use of hybrid designs, which apply Bayesian methods at interim analyses and a frequentist final analysis. This paper contributes to the evidence that fully Bayesian adaptive designs are a viable alternative, and investigates efficiency gains available through prior specification and participant allocation.

Methods

Using a Phase II, multi-arm, multi-stage platform trial as a case study, we converted the original hybrid design into a fully Bayesian framework by replacing the final frequentist test with a Bayesian posterior probability decision rule, calibrated to match the original significance levels. Interim efficacy and futility rules were retained. A simulation framework (100,000 synthetic trials per configuration) was used to compare both designs on four operating characteristics: assurance, Bayesian type I error, and their frequentist counterparts. Expected sample size (ESS) was estimated as a measure of efficiency. Informative truncated design priors governed data generation non-informative analysis priors governed within-trial decisions. Five participant allocation configurations were evaluated per arm.

Results

At zero prior variance, the fully Bayesian and hybrid designs yield identical frequentist performance, confirming that the Bayesian framework can exactly reproduce the operating characteristics of its hybrid counterpart when appropriately calibrated. As prior variance increases, the two frameworks diverge: Bayesian type I error declines sharply toward zero, while the hybrid design's frequentist type I error remains substantially higher. Assurance shows a modest initial dip before stabilising above its starting level. Balanced participant allocation configurations consistently produced the lowest ESS and highest assurance across both arm types and all variance levels.

Discussion

A fully Bayesian adaptive design can meet the frequentist operating characteristics that motivate the use of hybrid designs, directly addressing the central regulatory concern. Beyond this equivalence, it offers substantially better false positive control. Assurance provides a more realistic measure of trial success than frequentist power by formally incorporating prior uncertainty. Balanced participant allocation is a straightforward and actionable design optimisation that reduces ESS and increases assurance and should be considered as a default in multi-stage adaptive trials.

[28-P3-17] Adaptive Platform Trials Within a Cohort Study: A Novel Design with Application to PREMIS in Pregnancy Research

Presenter: Babak Choodari-Oskooei

Authors: Babak Choodari-Oskooei [1]; Emily Fox [2]; Lee Priest [2]

Affiliations: [1] UCL, London; [2] University of Birmingham, Birmingham

Introduction

Multi-arm multi-stage (MAMS) adaptive platform trials allow multiple interventions to be evaluated within a unified framework, but their implementation in population-based setting have not been explored adequately. Separately, Trials within Cohorts (TwiCs) leverage observational cohorts to improve trial feasibility, yet traditionally do not incorporate dynamic platform features. We introduce

PREMIS, a methodological extension that integrates platform trial architecture within an established pregnancy outcomes cohort.

Methods

PREMIS is an adaptive platform trial embedded within an established pregnancy outcome cohort study

(https://doi.org/10.1186/ISRCTN78792722), using the Trials-within-Cohorts (TwiCs) framework to evaluate interventions aimed at improving the chance of live birth among women with a history of miscarriage. The ongoing cohort provides the infrastructure for efficient participant identification, eligibility assessment, recruitment, and outcome collection. Eligible participants are randomised to intervention arms directly from within the cohort, enabling the cohort's epidemiological objectives and the platform trial's evaluative objectives to be achieved in parallel. The design retains key features of platform trials, including staged evaluation of multiple interventions, flexible addition/removal of intervention arms, and the potential use of adaptive decision rules. For the initial set of interventions, we describe the statistical framework underpinning a MAMS Selection design that incorporates an intermediate outcome to facilitate early treatment selection and accelerate decision-making.

Results

Embedding the platform within an existing cohort improves operational efficiency by reducing missing outcome data through routine data collection increasing recruitment yield by pre-identifying eligible individuals lowering trial start-up and infrastructure costs supporting dynamic intervention selection and sequential evaluation as well as providing a unified population denominator, enabling comparable control groups across trial stages.

Furthermore, pre-specification of a treatment selection rule reduces the maximum sample size by approximately 33% in our simulations. Finally, we compare the operating characteristics (i.e., family-wise type I error rate and power) of the proposed with alternative designs.

Discussion

PREMIS represents a methodological evolution of both platform trials and cohort-embedded RCTs. By integrating trial randomisation within a standing cohort, the design offers an efficient and pragmatic structure for evaluating multiple interventions across diverse maternal health contexts. This approach may generalise to other disease areas where established cohorts offer high-quality longitudinal data.

P4 - Causal Methods and Epidemiology

Poster area: P4

[28-P4-01] Bridging the Gap: Understanding attitudes towards causal inference methods in applied healthcare research

Presenter: Stephanie Riley

Authors: Stephanie Riley [1]; Molly Wells [1]; David Jenkins [1]; Matthew Sperrin [1]

Affiliations: [1] University of Manchester, Manchester

Introduction

Causal inference is increasingly recognised as a useful tool in healthcare research, enabling researchers to draw causal conclusions, even when patients are not randomised to treatments. Although many research questions are inherently causal, they are often framed in terms of "association" to avoid explicit causal claims. As such, uptake of causal inference methods remains limited. Guidance on communicating these approaches effectively is also lacking. To explore current attitudes, we designed a survey to identify barriers to understanding and implementing causal inference and to inform strategies for increasing its use in practice.

Methods

A descriptive cross-sectional study was conducted using an online questionnaire targeting researchers, analysts and methodologists working in healthcare or with healthcare data. The questionnaire collected information on participants' backgrounds, their experience with and views on causal inference, as well as their confidence in implementing causal methods. Closed-ended questions were summarised using descriptive statistics, and free-text responses were analysed to identify common themes.

Results

The survey ran for six weeks and received 122 responses, most from academia (71%). Participants represented a range of career stages including entry, mid, senior, and leadership roles, and a variety of disciplinary backgrounds. Overall, 89% were familiar with causal inference methods, and over half (53%) had applied them in practice. Among those who had not applied these methods, lack of confidence was the most frequently cited barrier. In terms of formal training, 61% reported training received through degree programmes. Analysis of free-text suggested that while introductory guidance was widely available, participants noted a lack of advanced resources, worked examples of causal studies, and standardised frameworks or best practice guidelines.

Discussion

Attitudes towards causal inference in healthcare research vary, ranging from high confidence in methodology and implementation to a lack of trust in the methods themselves and limited confidence in applying them in practice. Awareness of their potential and growing role in practice is clear, but hesitation remains. Practical guidance, targeted training, and wider availability of real-world case studies are required to close the gap between awareness and application.

[28-P4-02] Prioritization of adverse events related to integrase inhibitors and NNRTIs: a disproportionality analysis using data from the FAERS database

Presenter: Shuang Xia

Authors: Shuang Xia [1]; Gao-yang Wang [2]; Mayur Sarangdhar [3]; Yoshihiro Noguchi [4]; Ji-fang Zhou [2]

Affiliations: [1] Sorbonne University, Paris; [2] School of International Business, China Pharmaceutical University, Nanjing; [3] Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati; [4] Laboratory of Clinical Pharmacy, Gifu Pharmaceutical University, Gifu

Introduction

Background and objectives: Integrase strand transfer inhibitors (INSTIs) and NNRTIs are widely used in the treatment of HIV, but the real-world adverse events (AEs) profiles of these drugs remain inadequately characterized. We prioritized AEs reported with INSTIs and NNRTIs using data from the FDA Adverse Event Reporting System (FAERS).

Methods

Methods: Using FAERS data (from FDA approval to June 2024) for 10 drugs (cabotegravir, dolutegravir, raltegravir, elvitegravir, bictegravir, nevirapine, efavirenz, rilpivirine, doravirine and etravirine), we assessed AE clinical priority using a multi-criteria score. The score evaluated four criteria: clinical relevance, reporting proportion, case fatality rate and signal stability. We calculated reporting odds ratios (RORs) with 95% CIs signals required an AQ5 ROR025>1 after Bonferroni correction across [N] tested AEs per drug. Events related to medication errors or HIV-related conditions were excluded. A case-by-case assessment was conducted to evaluate the confounding factors from co-medications and indications

Results

Results: We found 4487/6362 (70.5%) were low priority (0-2 points) and 1849/6362 (29.1%) were moderate (3-5 points). Respectively, 2, 6, 9 and 15 marginally high-priority adverse pregnancy outcomes (5 points) were identified with dolutegravir, raltegravir, nevirapine and efavirenz 26 AEs were high clinical priorities (>5 points). After case-by-case assessment, five noteworthy AEs remained. These included efavirenz-related haemolytic anaemia, etravirine-related hepatic failure, dolutegravir-related foetal death, raltegravir-related drug reaction with eosinophilia and systemic symptoms, and raltegravir-related hepatic failure.

Discussion

Conclusions: This study provides a systematic framework for evaluating post-marketing AEs of INSTIs and NNRTIs using a semi-quantitative scoring system. Our findings identified five high-priority AEs that require clinical validation and further investigation.

[28-P4-03] Variable selection in regression models for observational studies. Possible pitfalls in interpreting results and possible biases

Presenter: Pascal Kemmerer

Authors: Pascal Kemmerer [1]; Emilio Gianicolo [1]; Andreas Stang [2, 3]

Affiliations: [1] University Medical Center of the Johannes Gutenberg University Mainz, Institute for Medical Biostatistics, Epidemiology and Informatics (IMBEI), Division of Epidemiology and Health Services Research, Mainz; [2] Institute of Medical Informatics, Biometry, and Epidemiology, University Hospital Essen, Essen; [3] Cancer Registry of North Rhine-Westphalia, Bochum

Introduction

In analyzing observational studies, variable selection can be based on either a predictive or a causal inference approach. However, choosing an approach that is ill-suited to the research question may contribute to incorrect results and/or incorrect interpretations of the results.

Methods

Based on a selective literature review, we illustrate how predictive and causal inference approaches imply different strategies of variable selection and different interpretations of regression coefficients in an observational study.

Results

Predictive approaches focus on discrimination and calibration to accurately predict an outcome. A regression model based on a predictive approach would contain any variable that improves prediction, regardless of causal relationships. In contrast, causal inference approaches rely on prior knowledge and focus on de-confounding to obtain an unbiased estimate of the effect of an exposure on an outcome. A regression model based on a causal inference approach would mainly contain variables that have been identified as exposure, confounder or outcome. From a causal inference perspective, we show possible misinterpretations and biases that can occur when predictive and causal approaches are not distinguished.

Discussion

An understanding of the differences between predictive and causal inference approaches is necessary in order to choose a variable selection strategy that is appropriate for the research question. Directed acyclic graphs may help to mitigate misinterpretations when answering causal questions.

[28-P4-04] To adjust or not to adjust? Evaluating BMI adjustment in associations between imaging-derived fat markers and cardiometabolic risk factors: A NORTA-based simulation study

Presenter: Elena Grune

Authors: Elena Grune [1, 2]; Anne-Laure Boulesteix [3]; Annette Peters [2]; Susanne Rospleszcz [1]

Affiliations: [1] Universitätsklinikum Freiburg, Klinik für Diagnostische und Interventionelle Radiologie, Freiburg im Breisgau; [2] Institute of Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health (GmbH), Neuherberg; [3] Institute for Medical Information Processing, Biometry, and Epidemiology, Faculty of Medicine, LMU Munich, Munich

Introduction

Population-based studies as KORA or NAKO demonstrated the feasibility of body composition assessment through magnetic resonance imaging (MRI). MRI-derived measures of abdominal and ectopic fat are increasingly analyzed in relation to cardiometabolic risks. Regression models used are often adjusted for body mass index (BMI), a conventional practice with an unclear rationale: BMI may act as a proxy for specific fat depots, reflect overall adiposity, or introduce redundancy due to correlation with imaging measures.

This simulation study aims to evaluate how adjustment for BMI influences regression estimates of associations between imaging-derived fat depots and cardiometabolic risk factors under varying realistic correlation structures.

Methods

Exposures including different fat depots, age, and BMI were simulated separately for women and men using the normal-to-anything (NORTA) approach. Marginal distributions and correlations were calibrated to empirical data from the KORA cohort. Continuous cardiometabolic outcomes comprised blood pressure, cholesterol, and HbA1c and were simulated jointly with exposures. Regression models are fitted with and without BMI adjustment and effects of varying correlation structures and sample sizes on fat beta coefficient estimates are evaluated. Performance measures include coefficient shift, standard error, and proportion of non-significant estimates.

Results

Preliminary simulations show that NORTA-generated data resemble empirical data. Marginals were chosen from parametric distributions (Weibull, Gamma, Lognormal, and Normal) using visual diagnostics and goodness-of-fit criteria (fitdistrplus, R). Using these marginals and empirical correlation matrix from KORA, variables were generated and truncated at observed minima and maxima (simdata, R). Simulated data reflected skewness, heterogeneous marginal distributions, and preserved the empirical correlation structure. Next, correlation structures will be systematically varied to reflect three hypothesized relationships between fat, BMI, and outcomes, while keeping other correlations at empirical values: 1) redundancy (high BMI-fat correlation) 2) independence (either fat-outcome or BMI-outcome correlation high, with complementary low) 3) balanced confounding (BMI and fat equally correlated with outcome). Simulations will be conducted across different sample sizes, regression models fitted, and performance measures compared.

Discussion

These simulations may highlight conditions under which BMI adjustment alters estimated associations and provide guidance for interpreting regression analyses of imaging-derived fat depots and cardiometabolic risk factors.

[28-P4-05] Selection bias can induce non-linear associations

Presenter: Apostolos Gkatzionis

Authors: Apostolos Gkatzionis [1]; Elinor Curnow [1]; Kate Tilling [1]

Affiliations: [1] University of Bristol, Bristol

Introduction

Selection bias is a common issue in epidemiological studies. Here, we focus on selection bias due to collider stratification: if the exposure (or a cause of the exposure) and the outcome (or a cause of the outcome) of a study affect selection into the study sample, restricting the analysis to the selected sample can bias estimates of the parameter of interest (e.g. regression coefficient or average causal effect). The nature of this bias, including whether it can affect the functional form of the exposure-outcome relationship, has not been well studied.

Methods

We show that a relationship between an exposure X and an outcome Y that is linear at a population level can appear to be non-linear in the selected sample due to selection bias. We establish conditions under which a non-linear X-Y relationship can be induced. We quantify, algebraically and by simulation, the magnitude of the non-linear association, considering settings with a linear, logistic, or Poisson regression analysis model and a logistic or log-additive selection model. We investigate whether common goodness-of-fit tests will detect non-linearity when applied to the selected sample. Finally, we explore the implications of our results for multiple imputation.

Results

Under a log-additive model with selection probability piS = exp left g1(x) + g2(x) y + g3(x) y2 right, a non-linear X-Y relationship will be induced if either g2(x) is not linear in x or g3(x) is not constant in x. Simulations suggest that this is also approximately the case under a logistic model. Likelihood ratio tests based on splines or fractional polynomials reject the null hypothesis of a linear X-Y relationship with high probability in samples of size n=1000-5000. Bias of the multiple imputation estimate does not increase if the imputation model is mis-specified based on observed relationships in the selected sample, provided the imputation model remains richer than the correctly specified model.

Discussion

The functional form of the association between an exposure and an outcome in the selected sample may be distorted due to selection bias. Therefore, findings of non-linear associations in studies that are susceptible to selection bias should be treated with caution. Careful consideration of the selection mechanism is important in such analyses.

[28-P4-06] A Comparative Study of Sensitivity Analysis Methods for Causal Inference in Observational Study

Presenter: Maliha Binte Alauddin

Authors: Maliha Binte Alauddin [1]; Mahbub A. H. M. Latif [1]

Affiliations: [1] Institute of Statistical Research and Training (ISRT), University of Dhaka, Dhaka

Introduction

The causal inference from observational data relies on the assumption of no unmeasured confounding, an assumption that cannot be tested empirically. Randomized controlled trials are regarded as the gold standard for causal inference because random assignment creates balance between groups, but they are often infeasible. As a result, researchers depend on observational studies, which face the unavoidable problem of confounding. Sensitivity analysis offers a way to assess how much hidden bias would be required to change the estimated effect and therefore helps to establish reliability of the strength of causal claims when this core assumption may fail.

Methods

This study presents a comparative study of three distinct sensitivity analysis frameworks. These include the model based approach of Rosenbaum and Rubin, the randomization based Rosenbaum Bounds, and the E-value, along with recent extensions for continuous outcomes. Through a set of Monte Carlo simulations, we assessed how these approaches perform under different levels of unmeasured confounding

Results

The results show that the adjusted estimator based on Rosenbaum and Rubin and the E-value can recover the true causal effect with small bias. At the same time, the findings indicate that Rosenbaum Bounds may not always give an unbiased estimate of robustness, especially when the true effect is small relative to the bias due to the unmeasured confounder.

Discussion

However, a detailed review of recent high impact causal research reveals omission of these robustness checks which leaves significant causal findings vulnerable to potential hidden biases. To address this gap, this study synthesizes the methodological strengths and limitations of each approach. The findings underscore that the use of formal sensitivity analysis is not merely a statistical refinement but a necessary condition for credible and reliable causal evidence.

[28-P4-07] Multiverse analysis for large-scale real-world evidence: Opportunities and challenges for robustness and transparency

Presenter: Mario Lawes

Authors: Mario Lawes [1]; Alexia Sampri [2, 3]; Ben Illigens [1, 4, 5]

Affiliations: [1] Data4Life, Berlin; [2] British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge; [3] Victor Phillip Dahdaleh Heart and Lung Research Institute, University of Cambridge, Cambridge; [4] Hasso Plattner Institut, Potsdam; [5] Dresden International University, Dresden

Introduction

Real-world evidence (RWE) from observational health data is vulnerable to the 'garden of forking paths', where defensible choices regarding, for example, cohort definition, covariate selection, and model specification can yield divergent results. This analytical flexibility challenges the reproducibility, interpretability, and credibility of observational findings. Multiverse analysis addresses this issue by systematically evaluating a range of plausible specifications and visualizing their impact on results. Despite growing use in the social sciences, it remains underused in epidemiology (for exceptions see 1,2).

Methods

We propose a structured framework for applying multiverse analysis in large-scale RWE studies, building on existing OHDSI methodology for cohort definition, covariate selection, and effect estimation. A central requirement is that all specifications must target the same underlying causal estimand, so that variation in results reflects analytical uncertainty rather than heterogeneity in the research question. The framework is implemented within Data2Evidence (D2E), an open-source platform for scalable, transparent RWE generation (3).

Results

We demonstrate how to execute multiverse analyses within D2E and interpret the resulting specification-curve visualizations. We further discuss key challenges, including the interpretation of large specification spaces and the effective communication of results to clinical audiences.

Discussion

Multiverse analysis can increase transparency and robustness in RWE studies, but only when implemented with care. Its value lies in quantifying and visualizing the effect of different analytical specifications and prompting reflection on analytic choices - not in replacing principled study design. The proposed guidelines support its responsible use in RWE contexts and its integration within the OHDSI infrastructure lowers barriers to adoption.

References

1. Levitt M, Zonta F, Ioannidis JPA. Excess death estimates from multiverse analysis in 2009-2021. Eur J Epidemiol. 2023 Nov38(11):1129-39. doi:10.1007/s10654-023-00998-2

2. Patel CJ, Burford B, Ioannidis JPA. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. J Clin Epidemiol. 2015 Sep68(9):1046-58. doi:10.1016/j.jclinepi.2015.05.029

3. Guerrero P, Ernebjerg M, Holst T, Weese D, DiBello H, Ibing S, et al. The AIR·MS data platform for artificial intelligence in healthcare. JAMIA Open. 2025 Nov 38(6):ooaf145. doi:10.1093/jamiaopen/ooaf145

[28-P4-08] Evaluation of the impact of outcome generation on method performance within the plasmode simulation framework

Presenter: John Tazare

Authors: John Tazare [1]; Marleen Bokern [1]; Christopher T. Rentsch [1]; Anna Schultze [1]; Elizabeth Williamson [1]

Affiliations: [1] London School of Hygiene and Tropical Medicine, London

Introduction

Plasmode simulations are used to evaluate and compare causal inference methods in real-world data settings. These simulations retain most of the structure of complex healthcare data while imposing known treatment effects, allowing method performance to be assessed under realistic conditions. However, plasmodes require specification of an outcome-generating model, and the extent to which conclusions depend on this choice is unclear. If outcome models favour certain analytical approaches, comparisons between methods may be misleading. We therefore aimed to investigate how the choice of outcome generation influences conclusions drawn from plasmode-based evaluations of causal inference methods.

Methods

Using a recent cohort study of inhaled corticosteroid (ICS) use in the UK Clinical Practice Research Datalink, we generated plasmodes containing 20,000 patients sampled from the underlying cohort (N= 78,336), with a true marginal risk difference of 1. Outcomes were simulated using logistic regression, generative additive models and random forests. The specification of these models ranged from minimal confounder adjustment sets to those additionally considering inputs from the high-dimensional propensity score (HDPS) algorithm. For the latter, we considered several representations of the covariate space: binary indicators, count frequences and principal components.

Results

Under these different outcome models, we compared the HDPS against machine learning alternatives for confounder selection including: least absolute shrinkage and selection operator, elastic net, and random forest. Comparative performance of methods was evaluated primarily by comparing bias in the marginal risk difference.

Discussion

The plasmode simulation framework is commonly used for evaluating causal inference methods, including to rank a set of candidate approaches in a given setting. However, empirical understanding of the impact of outcome generation choice on conclusions has not been previously investigated. Findings from our study will inform practical guidance on the conduct of plasmode simulations.

[28-P4-09] Federated Propensity Score Estimation

Presenter: Liane-Marina Meßmer, Nina Haug

Authors: Liane-Marina Meßmer [1]; Holger Ziekow [1]; Christoph Reich [1]; Nina Haug [2]

Affiliations: [1] Hochschule Furtwangen, Furtwangen; [2] iOMEDICO, Freiburg im Breisgau

Introduction

Real-world data complements randomized trials in oncology but requires causal methods to adjust for confounding. Propensity score (PS) weighting enables estimation of treatment effects, yet multi-center analyses are often constrained by data protection rules that prevent pooling patient-level data. Federated learning (FL) offers collaborative model training without sharing raw data however, it remains unclear how well federated PS estimation performs under realistic between-center heterogeneity.

Methods

We simulated PS estimation across three institutions under (i) homogeneous data, (ii) heterogeneous treatment prevalence across sites with equal sample sizes, and (iii) heterogeneous treatment prevalence with unequal sample sizes. A federated PS model (FL with periodic server aggregation) was compared to an otherwise identical centralized model trained on pooled data. We assessed PS quality via calibration and covariate balance after weighting, quantified by absolute standardized mean differences (mean and maximum). We additionally examined weight stability (extreme weights/effective sample size) to characterize practical impact on IPW-based effect estimation.

Results

Federated and centralized PS models were compared across increasing degrees of site heterogeneity to quantify when federation approximates pooled-data performance and when it leads to miscalibration, residual imbalance, or unstable weights.

Discussion

FL can enable multi-center causal analyses without sharing patient-level data, but its suitability for PS-based weighting depends on the heterogeneity structure and sample-size imbalance across sites. Our findings provide guidance on when federated PS estimation is reliable and which diagnostics should accompany federated causal analyses.

[28-P4-10] Scalable dual optimization functional propensity scores for estimating time-varying causal effects

Presenter: Francesca Ieva

Authors: Nicole Fontana [1, 2]; Simone Ciardulli [1]; Simone Vantini [1]; Francesca Ieva [1, 2]

Affiliations: [1] MOX, Department of Mathematics, Politecnico di Milano, Milano; [2] Health Data Science Research Centre, Human Technopole, Milan, Italy, Milano

Introduction

Causal effects of exposures often vary over the life course, with specific periods exerting stronger influence on health outcomes or revealing long-term impacts of risk factors. Capturing such time-varying causal effects is essential for understanding underlying mechanisms and informing effective

interventions. However, methodological tools for handling time-varying exposures remain limited in the causal inference literature.

Methods

We propose a functional propensity score (FPS) approach that extends propensity score methods to functional treatments using functional principal component analysis. Exposure trajectories are decomposed into principal components to capture the dominant modes of temporal variation. We build a functional propensity score model that balances exposure trajectories, creating weights based on confounders to ensure covariate adjustment, mimicking randomized assignment in the functional space.

Weight estimation is formulated as a smooth, unconstrained dual optimization problem and solved directly via Broyden-Fletcher-Goldfarb-Shanno quasi-Newton algorithm. Compared with existing approaches [1], our method eliminates the need for paramenter grid search and restrictive dimensionality assumptions while preserving exact balance in the reduced functional space.

Results

In simulation studies, the proposed method achieves near-zero bias, good coverage of causal effect trajectories across varying sample sizes and numbers of confounders, and reduced computational time compared with [1]. We further apply the method to estimate the time-varying causal effect of body mass index between ages 50 and 65 on the risk of type 2 diabetes using UK Biobank data.

Discussion

This framework integrates functional data analysis with causal inference to better characterize how risk factors evolve over time and affect health outcomes. It enables scalable and interpretable causal estimation for functional treatments with reduced reliance on tuning procedures. By identifying critical exposure periods, the approach provides clinically relevant insights that can inform targeted prevention strategies.

[1]X. Zhang, W. Xue, and Q. Wang. Covariate balancing functional propensity score for functional treatments in cross-sectional observational studies. Computational Statis- tics & Data Analysis, 163(C), 2021. doi: 10.1016/j.csda.2021.10730.

[28-P4-11] Causal Inference Using Propensity Score Matching: Effect of Tracheostomy Timing on Outcomes in TBI

Presenter: Vineet Kumar Kamal

Authors: Vineet Kumar Kamal [1]

Affiliations: [1] All India Institute of Medical Sciences (AIIMS), Nadia

Introduction

Traumatic brain injury (TBI) patients admitted to neurocritical care often require airway protection and prolonged mechanical ventilation. Tracheostomy is commonly performed in moderate-to-severe TBI however, the optimal timing remains controversial. Observational studies are prone to selection bias, limiting causal interpretation. This study aimed to estimate the causal effect of tracheostomy timing (early vs late) on functional outcomes, in-hospital mortality, and length of hospital stay in TBI patients.

Methods

We conducted a retrospective cohort study of 1,782 TBI patients admitted to a public, tertiary care level-1 trauma centre in India, of whom 529 underwent tracheostomy. Early tracheostomy was defined as ≤7 days post-admission and late as >7 days. Outcomes included: (1) six-month functional outcome (favourable vs unfavourable, based on Glasgow Outcome Score), (2) in-hospital mortality, and (3) length of hospital stay.

To address confounding inherent in observational data, propensity score matching (PSM) was employed. Propensity scores were estimated using multivariable logistic regression including clinically relevant covariates. Patients were matched using 1:1 nearest neighbour matching with a caliper of 0.05, ensuring covariate balance between groups. The average treatment effect on the treated (ATET) was estimated. Adjusted logistic and linear regression models were used to assess binary and continuous outcomes, respectively. Analyses were performed using Stata version 18

Results

Before matching, early tracheostomy was not significantly associated with length of hospital stay (β = -0.42 95% CI: -0.88 to 0.02), and results remained similar after matching (β = -0.38 95% CI: -0.79 to 0.01). No significant association was observed with favourable six-month functional outcome in regression models (OR reduced from 1.38 [95% CI: 0.63-3.02] to 1.18 [95% CI: 0.50-2.78] after matching).

However, causal effect estimation using PSM demonstrated a significant improvement in favourable six-month functional outcome (ATET: 0.15 95% CI: 0.07-0.24 p = 0.001) associated with early tracheostomy.

For in-hospital mortality, no significant association was observed before (OR = 1.99 95% CI: 0.92-4.27) or after matching (OR = 1.62 95% CI: 0.73-3.60), and ATET analysis confirmed no causal effect (ATET: 0.002 95% CI: -0.053 to 0.058 p = 0.938).

Discussion

Early tracheostomy was not associated with reduced hospital stay or in-hospital mortality in TBI patients. However, after accounting for selection bias using propensity score methods, early tracheostomy demonstrated a significant causal effect on improving six-month functional outcomes. These findings highlight the importance of causal inference approaches in observational neurocritical care research.

[28-P4-12] Causal Inference with Competing Risks Using Propensity Score Methods

Presenter: Mahin Tatari

Authors: Mahin Tatari [1]; Stefano Rosato [2]; Paola D'Errigo [2]; Giovanna Jona Lasinio [1]

Affiliations: [1] Department of Statistical Science, Sapienza University of Rome, Rome; [2] Istituto Superiore di Sanità, Rome

Introduction

Competing risks are common in survival analysis and the causal interpretation of classical competing risk estimands is unclear, particularly when competing events lie on the causal pathway and may act as mediators. This study extends the counterfactual framework for competing risks by incorporating propensity score (PS) methods to estimate the controlled direct and total effects within a causal mediation setting.

Methods

We analyzed data from 5,621 patients with severe aortic stenosis who underwent either transcatheter aortic valve replacement (TAVR) or surgical aortic valve replacement (SAVR). We used PS-based stabilized inverse probability of treatment weights (IPTW) to adjust for confounding treatment. To address competing event-outcome confounding, we constructed sequential stabilized weights from the treatment and mediator PS models. Weighted Cox proportional hazards models estimated total effects (TE), controlled direct effects (CDE), and both cause-specific and subdistribution hazards. Counterfactual 5-year risks were derived from weighted survival curves, and we calculated risk differences (RD) and risk ratios (RR).

Results

During the 5-year follow-up, The estimated TE among patients receiving TAVR on the risk of heart failure was slightly lower, not statistically significant (HR = 0.92, 95% CI: 0.77-1.11). The CDE showed no evidence of a direct association with heart failure risk (HR = 0.98, 95% CI: 0.81-1.20). The corresponding counterfactual 5-year risks yielded minimal absolute and relative differences in risk. The marginal hazard of heart failure under elimination of the competing event was (HR = 0.98, 95% CI: 0.80-1.20, p = 0.85). The cause-specific hazard for only heart failure (deaths as censoring) and for all-cause mortality corresponding to the TE was 0.88 (95% CI: 0.73-1.06 p = 0.18) and 1.22 (95% CI: 0.98-1.51 p = 0.071) respectively.

Discussion

Within counterfactual framework, classical estimands from the competing risks literature can be interpreted as causal contrasts of total and direct effects in failure-time settings where competing events may alter the probability of observing the outcome of interest. By adopting this causal perspective, the interpretation of commonly used survival estimands and allows the decomposition of treatment effects into components operating through different causal mechanisms.

[28-P4-13] Causal Mediation Analysis with Longitudinal Data: Revisiting the Difference Method

Presenter: Robin Denz

Authors: Robin Denz [1]; Nina Timmesfeld [1]

Affiliations: [1] Ruhr-Universität Bochum, Abteilung für Medizinische Informatik, Biometrie und Epidemiologie, Bochum

Introduction

Mediation analysis is a powerful tool to investigate mechanisms by which a treatment causally influences an outcome. Its main goal is to decompose total treatment effects into direct and indirect components, with respect to a mediating variable, and thus enables explanations of total treatment effects in a formal statistical framework. However, applying such analyses to settings with complex longitudinal data remains challenging. Existing methods are statistically complex, computationally intensive, and rarely available in user-friendly software. The widely known difference method offers a simple alternative. This method utilizes two models for the outcome: one to estimate the direct effect and one for the total effect. The difference of the obtained beta coefficients is used to estimate the indirect effect. However, the performance of this method in settings with longitudinal data has not been systematically evaluated.

Methods

We conducted a comprehensive Monte-Carlo simulation study to investigate the performance of the difference method in settings with a time-fixed treatment, a time-to-event outcome, time-dependent mediators and time-dependent confounders. Using Cox proportional hazards, Aalen additive hazards, and accelerated failure time (AFT) models with time-varying covariates, the difference method was compared across different data generation processes with time-dependent mediators and confounders, focusing on bias in estimated indirect effects. The parametric mediational g-formula was used as benchmark comparator.

Results

If correctly specified, the Aalen model based difference method produced unbiased estimates in the absence of a time-dependent confounder that was directly caused by the treatment. Similar results were obtained when using the Cox model based difference with rare outcomes, but not with common outcomes. The AFT model based difference method showed small bias in almost all scenarios, due to collapsibility issues. Only the parametric mediational g-formula was unbiased in all scenarios.

Discussion

In contrast to specialized methods, the difference method requires additional, often unrealistic, assumptions, such as the absence of a direct causal relationship of the treatment on time-dependent confounders. If those assumptions hold, however, it may be used as a simple and efficient alternative.

[28-P4-14] Causal Discovery and Bayesian Effect Estimation under Latent Confounding: One-Year Impacts of a Diabetic Nephropathy Prevention Program

Presenter: Yasutaka Hasegawa

Authors: Yasutaka Hasegawa [1]; Shuzo Ishizaka [1]; Takanobu Osaki [1]; Hideyuki Ban [1]; Takayuki Arai [2]; Tsutomu Kikuchi [2]

Affiliations: [1] Hitachi, Ltd. Research & Development Group, Tokyo; [2] Hitachi Health Insurance Society, Tokyo

Introduction

Preventive programs evaluated with routine health checkup and claims data are prone to both measured and unmeasured confounding. We applied a causal discovery-based approach to estimate one-year effects of a diabetes nephropathy progression-prevention program.

Methods

We used a two-stage strategy. First, Linear Non-Gaussian Acyclic Model (LiNGAM) was repeatedly applied to the program participation indicator, 37 candidate covariates, and each outcome to extract outcome-specific measured confounders. Second, Bayesian mixed-effects LiNGAM (BM-LiNGAM) accommodated latent confounding by transforming models with an unobserved confounder into equivalent mixture representations without latent variables, comparing causal directions via log marginal likelihood, and estimating average treatment effects (ATE) by MCMC. The analysis included 857 eligible individuals with one-year follow-up checkup data (participants n=66 nonparticipants n=791). Outcomes were 10 one-year clinical measures: body weight, waist circumference, systolic/diastolic blood pressure, triglycerides, HDL-C, LDL-C, HbA1c, fasting plasma glucose, and eGFR.

Results

After adjustment for measured and latent confounding, program participation showed posterior evidence of improvement in anthropometrics: weight -1.367 kg and waist -1.379 cm (each with ≥95% posterior probability of improvement). Estimated effects on blood pressure (SBP -1.198 mmHg DBP -0.486 mmHg), lipids (triglycerides -4.301 mg/dL HDL-C +0.367 mg/dL LDL-C +6.881 mg/dL), and glycemia (HbA1c -0.021% fasting glucose +0.915 mg/dL) were small and did not provide strong evidence of improvement. For kidney function, adjusting for measured confounding only suggested an eGFR increase of 4.553 ml/min/1.73m^2 (95% Bayesian credible interval 0.036 to 9.150 ≥95% posterior probability of improvement), whereas additional adjustment for latent confounding attenuated the estimate to 4.493 (-0.168 to 8.995) without ≥95% posterior evidence of improvement. In a subgroup of participants achieving ≥2 kg weight loss (n=23), the eGFR difference versus nonparticipants was 8.988 with ≥95% posterior probability of improvement.

Discussion

Combining confounder discovery with Bayesian causal modeling can quantify uncertainty over causal direction and effects in real-world prevention programs. The eGFR results illustrate how conclusions

can change when latent confounding is considered and suggest that weight reduction may be a key pathway to renal benefit.

[28-P4-15] Using causal network to compare patient populations across clinical studies and real-world data sources

Presenter: Svetlana Le Ralle

Authors: Svetlana Le Ralle [1]; David Pau [2]; Claire Castagné [2]; Markus Elze [3]; Sarah Zohar [4]; Laurent-Philippe Albou [5]

Affiliations: [1] F. Hoffmann-La Roche AG, Fontenay-lès-Briis; [2] Roche SAS, Boulogne-Billancourt, France, Boulogne-Billancourt; [3] F. Hoffmann-La Roche AG, Basel; [4] Inserm, Université Paris Cité, Inria, UMR 1346, HeKA, Paris; [5] Institut Roche, Boulogne-Billancourt, France, Boulogne-Billancourt

Introduction

Limited data availability, particularly in rare and paediatric indications, remains a significant challenge for clinical research. Comparing Randomised Clinical Trials (RCTs) and real-world data (RWD) helps identify shared patterns and biases across studies, data sources, and therapeutic areas. To support similarity assessment, heterogeneity quantification, and the feasibility of information borrowing, we developed a causal network-based framework to characterise and compare population structure across clinical studies and data sources.

Methods

Study-specific causal networks were inferred using MIIC (Multivariate Information-based Inductive Causation) (1, 2, 3), an information-theoretic causal discovery approach. MIIC learns a graph representing direct dependencies between variables, using the inferred structure to derive a concise approximation of the population's multivariate distribution. Robustness was assessed through bootstrap resampling, split-half reproducibility analyses, and null-model permutation testing. Bootstrap analyses quantified edge stability and node degree variability across resamples. Split-half analyses evaluated reproducibility using, for instance, structural Hamming distance and Jaccard similarity. Null datasets were generated to determine whether the learned graphs reflected a genuine multivariate structure rather than random patterns. Cross-study comparisons, like Graph Edit Distance or Node/Edge Overlap Measures, are used to compare study-specific networks.

Results

We assembled and harmonised variables from 15 RCTs and 5 RWD studies in 4 indications (including rare and paediatric diseases) within the INVENTS project (4), encompassing 10,000 patients. We generated study-specific causal networks with robustness metrics and cross-study comparison metrics. This approach characterised the causal relationships within each study to create a causal signature, used to identify related studies that may be suitable for information borrowing.

Discussion

This framework provides a principled and extensible approach to compare patient populations through multivariate dependency structures rather than isolated variables. It facilitates systematic comparisons between RCT and RWD data sources, providing a foundation for determining when structural consistency justifies borrowing information across sources.

References:

1) https://doi.org/10.1371/journal.pcbi.1005662

2) 10.1371/journal.pcbi.1007866

3) https://doi.org/10.1038/s41746-025-01431-6

4) https://invents-he.eu/

INVENTS has received funding from the European Union's Horizon Europe Research and Innovation program under grant agreement 101136365, the Swiss State Secretariat for Education, Research and Innovation (SERI) and the UKRI Innovative UK under their Horizon Europe Guarantee scheme.

[28-P4-16] Causal Inference of Trial Data for Decision-analytic Modeling: Applying Inverse-probability-of-censoring Weighting to RECETAS Trial Data

Presenter: Ursula Rochau

Authors: Ursula Rochau [1]; Veronika Papon [1]; Marjan Arvandi [1]; Lára R. Hallsson [1]; Beate Jahn [1]; Gaby Sroczynski [1]; Jill S. Litt [2, 3, 4]; Laura Coll-Planas [5, 6]; Montse Masó-Aguado [5, 6]; Jùlia Santamaria Navarro [1]; Igor Kuchin [1]; Sibylle Puntscher [1]; Uwe Siebert [1, 7, 8]

Affiliations: [1] Institute of Public Health, Medical Decision Making and Health Technology Assessment, Department of Public Health, Health Services Research and Health Technology Assessment, UMIT TIROL - University for Health Sciences and Technology, Hall in Tirol; [2] Institute for Global Health (ISGlobal), Barcelona Biomedical Research Park (PRBB) Doctor Aiguader, Barcelona; [3] CIBER Epidemiología y Salud Pública (CIBERESP), Madrid; [4] Universitat Pompeu Fabra (UPF), Barcelona; [5] Research Group On Methodology, Methods, Models and Outcomes of Health and Social Sciences (M3O), Faculty of Health Sciences and Welfare, Centre for Health and Social Care Research (CESS), University o, Barcelona; [6] Institute for Research and Innovation in Life Sciences and Health in Central Catalonia (IRIS-CC), Vic; [7] Center for Health Decision Science, Departments of Epidemiology and Health Policy & Management, Harvard Chan School of Public Health, Boston; [8] Institute for Technology Assessment and Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston

Introduction

Pragmatic randomized controlled trials (RCT) often suffer from missing data, loss to follow-up and differential drop-out impacting validity and generalizability. The pragmatic RECETAS (EU Horizon2020 project) Friends-in-Nature-Barcelona (FiN-BCN) trial evaluated nature-based social prescribing (NBSP) against usual care (UC). We use data from FiN-BCN to show how multivariate imputation by chained

equations (MICE), inverse-probability-of-censoring weighting (IPCW) to correct for differential drop-out, and adjusting for potential baseline confounding impact results of a decision-analytic modeling-based analysis.

Methods

We developed a Markov state-transition model including a combination of loneliness health states and loneliness-related diseases. Outcomes were quality-adjusted life years (QALYs) and loneliness-free years (LFYs). The model was populated, amongst others, with FiN-BCN trial (n=320) data. Due to missing values, we applied MICE. Drop-out rate in the control arm (n=44) was twice as high as in the intervention arm. A high proportion of people dropping out disagreed with assignment to control arm indicating differential drop-out leading to post-baseline selection bias. Therefore, we applied IPCW. As baseline characteristics showed imbalances, we additionally adjusted in the analyses for sex and flourishing scale.

Results

Populating our decision-analytic model with data from complete cases (n=170) only, resulted in 15.79 QALYs (15.07 LFYs) for a Spanish cohort of 61-year-old people over a lifelong time horizon for NBSP. Incremental QALYs were 0.55 (3.78 LFYs) for NBSP compared to UC. After imputing missing values, our decision-analytic model resulted in 15.30 QALYs (11.79 LFYs) for NBSP leading to incremental QALYs of 0.47 (3.12 LFYs). Additionally applying IPCW, led to 15.29 QALYs (11.60 LFYs) for NBSP. Incremental QALYs increased to 0.53 (3.58 LFYs) in favor of NBSP. Adjusting additionally for potential baseline confounding resulted in 15.29 QALYs (11.59 LFYs) for NBSP leading to incremental QALYs of 0.52 (3.46 LFYs).

Discussion

Exploring different options of causal analyses of pragmatic RCTs is important as modeled differences in treatment strategies can vary substantially. In our case, performing MICE compared to complete case analysis decreased incremental QALYs by 15% whereas adding IPCW to MICE with or without baseline adjustments results in almost identical incremental QALYs (<6% difference) compared to complete case analysis for NBSP versus UC.

P5 - Time-to-event and Longitudinal Methods

Poster area: P5

[28-P5-01] A state-machine based implementation of dynamic longitudinal study workflows in REDCap: Application to child wish, pregnancy, and postpartum surveillance in inflammatory rheumatic diseases

Presenter: Marius Schwarzwäller

Authors: Marius Schwarzwäller [1]; Günther Rezniczek [2]; Yvette Meißner [1, 3]; Tanja Ellmann [1]; Rebecca Fischer-Betz [4, 5]; Isabell Haase [6]; Jutta G. Richter [4, 5]; Tatjana Rudi [1]; Matthias Schneider [4, 5]; Susanna Späthling-Mestekemper [7]; Angela Zink [1]; Adrian Richter [1, 8]; Anja Strangfeld [1, 9]

Affiliations: [1] Epidemiology and Health Services Research, German Rheumatology Research Center, Berlin; [2] University Hospital of the Ruhr University Bochum, Herne; [3] Institute for Social Medicine, Epidemiology and Health Economics, Charité - University Medicine Berlin, Berlin; [4] Department for Rheumatology, University Hospital Düsseldorf, Medical Faculty of Heinrich-Heine-University, Düsseldorf; [5] Hiller Research Center, University Hospital Düsseldorf, Medical Faculty of Heinrich-Heine-University, Düsseldorf; [6] Division of Rheumatology and Systemic Inflammatory Diseases, III. Department of Medicine, University Medical Center Hamburg-Eppendorf, Hamburg; [7] Rheumatology Practice München-Pasing, München; [8] Department SHIP-KEF: Clinical-Epidemiological Research, University Medical Center Greifswald, Institute for Community Medicine, Greifswald; [9] Department of Rheumatology and Clinical Immunology, Charité University Medicine, Berlin

Introduction

Disease registries are frequently structured around fixed follow-up intervals and uniform outcome definitions. The Rhekiss registry, established in 2015, captures data on women with inflammatory rheumatic diseases (IRD) across different reproductive states, i.e. "child wish", "pregnancy" and "postpartum period" [1]. The current electronic data capture (EDC) solution follows a predefined structure and is associated with increasing operational costs, posing challenges for sustainable investigator-initiated academic studies. Further limitations of the current solution, including limited flexibility, restricted data control and dependency on external providers, are addressed by transitioning Rhekiss to REDCap.

Methods

A state-machine framework was designed to model dynamic patient trajectories. A set of clinically relevant and administrative states and respective entry criteria were defined. Additionally, mutually non-exclusive reproductive states were used to develop a priority hierarchy that ensures deterministic behavior of the EDC-system including the choice of state-dependent electronic case-reporting-forms (eCRFs). Built-in REDCap functionality was configured to control follow-up timing and eCRF selection. Furthermore, questionnaire wording is dynamically adapted to the patient's current situation.

Results

The framework was operationalized using REDCap events, calculated fields, branching and form display logic, and the survey queue. At each follow-up, the participant's current reproductive status is dynamically assessed using questionnaire items (checkboxes) that determine whether (1) a previously reported pregnancy is ongoing, (2) a pregnancy has ended since the last visit, (3) a current pregnancy has not yet been reported, and (4) there is a desire to conceive. Overlapping states are resolved according to a predefined priority hierarchy, with pregnancy taking precedence over family

planning. Calculated date fields determine the time to the next visit, while conditional branching displays context-specific response options to reduce input errors. Data entry triggers and automated notifications deliver a largely automated workflow.

Discussion

Child wish and pregnancy registries for women with IRD require more comprehensive data collection than other disease cohorts, including domains such as maternal health and treatment exposure. Flexible systems such as REDCap with individualized follow-up are essential due to varying pre-conception, pregnancy and postpartum courses. Migration of Rhekiss to REDCap is ongoing, with first results expected at the 2026 GMDS conference.

References

[1] PMID: 39138926

[28-P5-02] Calculating power and sample size for longitudinal mixed models with a new Stata program 'mixedpower'

Presenter: Matthew Burnell

Authors: Matthew Burnell [1]; James Carpenter [2]; Ian White [3]

Affiliations: [1] University College London, London; [2] UCL Innovative CLinical Trial Unit, London; [3] UCL, London

Introduction

Power and sample size calculations for clinical trials with longitudinal continuous outcomes measures are typically performed using simulation, with a perception that analytic solutions are either complex or intractable. However, there is a straightforward general approach using matrix algebra that is perhaps not widely appreciated and has been under-utilised.

Methods

We demonstrate the approach with a simple example where the treatment effect is a proportionate shift in mean slope. We then explain how the principle can be easily extended to any desired treatment effect specification and any required covariance structure. We also demonstrate how it is possible to ascertain the impact on power and resulting bias from incorrectly assuming a specific treatment effect, such as a proportionate slope change. Other essential requirements of a realistic trial, such as dropout, staggered recruitment and unequal allocation ratios can easily be incorporated.

Results

We present a new Stata command 'mixedpower' which can accommodate all the complexities described above. It allows for many different parameterisations of both control and treatment arms, including user-specified versions. One may elect to define complex level 1 error patterns such as auto-regressive, Toeplitz or even unstructured residuals across time, and both error structure and

random effects terms may differ by treatment group if desired. Crucially, all of the variance parameters may be entered either manually but also automatically, extracted from a fitted model, saving both time and potential mistakes. Companion programs are introduced that allow for easy identification of incomplete follow-up patterns due to dropout and staggered recruitment that calculate the amount of follow-up needed for required information fractions and perform power calculations for advanced models such as multivariate mixed models when there is more than one outcome.

Discussion

Power calculations for linear mixed models using simulation are time consuming, include uncertainty due to Monte Carlo error, and limit the number of scenarios that can be realistically considered. We have shown that these calculations can be easily and instantaneously performed using our new command 'mixedpower', demonstrating how this may be done even for complex models and set-ups necessary for the confident planning of a future clinical trial.

[28-P5-03] Impact of summer heat on emergency ambulance calls in Gelsenkirchen-a longitudinal study

Presenter: Marina Arendt, Moenes Sboui

Authors: Marina Arendt [1]; Laura Anderle [2]; Maren Berendonck [3]; Lise Don [3]; Peter Enste [4]; Theresa Weddeling [3]; Jenny Wielga [4]; Katja Zeume [2]; Moenes Sboui [2]

Affiliations: [1] Westfälische Hochschule Gelsenkirchen, Bocholt; [2] Westfälische Hochschule Gelsenkirchen, Gelsenkirchen; [3] Stadt Gelsenkirchen - Referat Gesundheit, Gelsenkirchen; [4] Institut Arbeit und Technik, Westfälische Hochschule Gelsenkirchen, Bocholt, Reckinghausen, Gelsenkirchen

Introduction

Heatwaves during summer present an increasing health risk, particularly for acute conditions such as cardiovascular disease, acute kidney failure, and heat stroke. Understanding, modelling and predicting the impact of high summer heat on health outcomes, specifically concerning emergency medicine is crucial for establishing preventive measures, as well as for emergency resource planning.

The core research questions are: Are there associations between heat and specific severe health outcomes? Does the frequency of emergencies during heatwaves differ from that in non-heat summers? Can time series analyses be used to predict emergency service needs with respect to heat conditions?

Methods

The study includes all emergency ambulance operations in Gelsenkirchen from 2010 to 2025. Data are linked with outside air temperature, building types in the area of the call, and social indicators of the affected city district. Adjustments are made for the relative population density in the respective geographical entity. Results are compared with previous work (e.g. Steul et al., 2018).

Results

The datasets comprise between 18,000 and more than 30,000 ambulance deployments per year. In addition to date and time, information on the initial and final diagnoses as well as deployment coordinates and prevailing weather conditions are available. This comprehensive database enables a differentiated analysis of the relationship between heat and emergency events.

Discussion

The presented work is still ongoing the current results section is only preliminary. The data set currently in analysis provides extraordinary insights due to its completeness, its geospatial information, and its time span of 15 years, including several heatwaves (2015, 2018, 2019). It should be noted that the data are based on the initial diagnoses made by emergency personnel. This study provides an important contribution to the understanding of heat-related emergency events and to optimizing emergency response planning in the context of climate change.

References:

Steul, Latasch, Jung, Heudorf (2018): Morbidität durch Hitze - eine Analyse der Krankenhauseinweisungen per Rettungseinsatz während einer Hitzewelle 2015 in Frankfurt/Main. Gesundheitswesen 80: 153-159.

[28-P5-04] Identifiability of Latent Trajectories with Neural ODEs in Small Data Regimes

Presenter: Timo Rachel

Authors: Timo Rachel [1]; Hanning Yang [1]; Tim Litwin [1]; Meropi Karakioulaki [2]; Cristina Has [2]; Moritz Hess [1]

Affiliations: [1] Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center - University of Freiburg, Freiburg; [2] Department of Dermatology, Medical Faculty and Medical Center - University of Freiburg, Freiburg

Introduction

Inflammatory biomarkers such as C-reactive protein (CRP) and interleukin-6 (IL-6) provide crucial insight into disease activity and treatment response. In inflammatory diseases where these biomarkers play a central role, clinical routine involves dense baseline measurements, while subsequent follow-up assessments - particularly in rare diseases such as epidermolysis bullosa (EB) - are sparse and irregularly spaced. This distorted observation process prevents reliable inference of the underlying disease course over time, even if there is information about the relationship between the biomarkers. To overcome the information-sparse small-data regime we augment the limited measurements with a Neural Ordinary Differential Equation (Neural ODE), which learns a continuous-time latent representation capable of reconstructing distinct patient trajectories despite the sparse and irregular sampling.

Methods

Using an ODE based on prior expert knowledge as ground truth we generate realistic synthetic biomarker trajectories for several sub-groups with different initial conditions and dynamic parameters. To avoid bias toward a single dynamical model, synthetic data used to train the Neural ODE was sampled from prior distributions over ODE parameters, covering a realistic spectrum of inflammatory courses. A joint autoencoder-Neural ODE architecture is trained: the encoder maps each biomarker vector into a latent representation, while a latent-space Neural ODE regularizes the temporal evolution of these latent states.

Results

We demonstrate under which conditions of the observation process Neural ODEs can still reliably reconstruct latent patient trajectories and separate subgroups in the latent space. Our simulations examine how reconstruction error, subgroup separability, and trajectory stability change as measurement sparsity and irregularity increase. Furthermore, we identify critical thresholds for sampling density and missingness patterns beyond which identifiability of the underlying dynamics breaks down.

Discussion

Using synthetic trajectories generated from a realistic ODE based on prior expert knowledge as training data, our joint autoencoder-Neural ODE architecture learns nonlinear dynamics and flexible latent representations. The combination of realistic data and nonlinear continuous-time modeling offers advantages over other methods proposed for irregular and noisy time series like linear ODEs, RNNs like LSTM, and latent ODE variants when handling the severe sparsity typical of rare disease biomarker data.

[28-P5-05] Continuous versus Discrete Modeling of Disease Progression: A Framework for Detecting True and Model-Induced Structure in Multiple Sclerosis

Presenter: Clemens Schächter

Authors: Susanne Weber [1]; Clemens Schächter [1]; Sibylle Frase [2]; Emiel Van den Hoven [2]; Heinz Wiendl [2]; Harald Binder [1]

Affiliations: [1] Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center - University of Freiburg, Germany, Freiburg; [2] Klinik für Neurologie und Neurophysiologie - Universitätsklinikum Freiburg, Freiburg

Introduction

Disease progression in chronic conditions such as multiple sclerosis (MS) is commonly described in discrete stages, although it may evolve gradually. This raises two questions: (i) whether progression is continuous or stepwise, and (ii) modeling issues when imposing discreteness. For developing a framework to assess discreteness of underlying structure and to quantify model-induced artifacts,

we systematically perturb data to create artificial discreteness. We fit discrete and continuous models and derive quantities indicating discreteness.

Methods

As a motivating application we consider multivariate longitudinal MS data with sparse, irregular follow-up and heterogeneous markers, summarized into latent scores. For inducing discreteness, we flagged tail observations in distributions and trajectories, subsequently restricting modeling to extreme individuals. Modeling results were contrasted to repeated random subsets of the original data.

As candidate approaches for modeling disease dynamics, a linear Gaussian state-space model (SSM continuous approach) was considered and a hidden Markov model (HMM discrete approach), complemented by a variational autoencoder (VAE) capable of both structures.

To derive quantities that might indicate discreteness for SSM, we evaluated trajectory geometry (smoothness, spread, sensitivity) and distributional properties (e.g., skewness) capturing symmetry changes under data perturbations. For HMM we developed a state-mapping approach and assessed stability via label switching.

Results

Visualization of VAE representations and corresponding separation metrics showed that the proposed selection approach induces artificial discreteness. In the original data, the SSM yielded smooth patient-specific trajectories, consistent with gradual progression. In the artificially discrete data, trajectory summary statistics changed. Specifically, trajectory features (slopes, step changes) became more symmetric under extreme subgroup restriction, consistent with reduced dynamics.

For the HMM, we were able to map states across fits and found correspondence with VAE/SSM. Artificially induced discreteness affected stability and mapping, enabling quantifying deviations from discreteness.

Discussion

Continuous SSMs and VAEs provide flexible representations of gradual dynamics, while discrete models yield sampling-sensitive summaries. All models provided promising quantities indicating discreteness in data. We emphasize assessing robustness and model-induced structure.

Data modification can induce apparent discreteness, providing a basis for a framework enabling detection of true and model-induced discreteness.

More generally, the proposed approach improves characterization of longitudinal patterns and may thus advance biomedical understanding.

[28-P5-06] Covariate-stratified disease course mapping: an extension of the Leaspy framework

Presenter: Sophie Tezenas du Montcel

Authors: Gabrielle Casimiro [1]; Sofia Kaisaridi [1]; Stéphanie Allassonnière [2]; Sophie Tezenas du Montcel [1]

Affiliations: [1] Sorbonne University, Paris Brain Institute (ICM), INSERM, INRIA, CNRS, APHP, Paris; [2] Université Paris Cité, Inria, Inserm, HeKA, Paris

Introduction

Disease progression models are promising tools for analysing longitudinal data presenting multiple modalities. Such models can be used to estimate a long-term disease progression and to reconstruct individual trajectories, while accounting for the variability between patients and features. Among these, the Disease Course Mapping (DCM) model, a nonlinear mixed-effects model implemented in the open-source Leaspy software, has demonstrated strong performance across a range of neurodegenerative diseases. Its key strengths lie in its ability to align heterogeneous follow-up timelines through individual temporal reparameterization, and to characterize differences in the relative ordering of feature degradation across patients through spatially structured random effects. However, the current framework does not allow the integration of patient-level covariates, limiting its applicability for clinical subgroup analysis.

Methods

We propose an extension of the logistic formulation of the DCM model to incorporate time-fixed covariates. In this formulation, the population-level trajectory for each feature is fully determined by three fixed effects: a reference time t0, a reference value g, and a velocity v. Linear covariate effects are introduced on each of these parameters, yielding a continuum of trajectories indexed by covariate values, around which individual random effects capture residual inter-individual variability. The covariate effect parameters are endowed with Gaussian priors and treated as latent fixed effects, jointly estimated with all other model parameters via a Monte Carlo Markov Chain Stochastic Approximation Expectation-Maximisation (MCMC-SAEM) algorithm. Structural identifiability of the extended model is established provided the covariate design matrix has full column rank.

Results

The model was validated on synthetic data with known covariate effects, chosen to reflect effects of covariates such as sex or APOE genotype on Alzheimer's disease progression. We show that the model successfully recovers ground truth parameters.

Discussion

This extension enables covariate-stratified disease course mapping while preserving the core properties of the Leaspy framework. Beyond group-level inference, the joint estimation of covariate effects is expected to yield a cleaner decomposition of individual variability: in the absence of an explicit covariate term, individual random effects partially absorb systematic group-level variability. The proposed model addresses this limitation, opening direct applications to clinical subgroup characterisation and covariate-adjusted patient staging in neurodegenerative diseases.

[28-P5-07] Prediction of latent disease status in a multivariate longitudinal model on Duchenne Muscular Dystrophy patients

Presenter: Chiara Degan

Authors: Chiara Degan [1]; Roula Tsonaka [1]; Pietro Spitali [2]; Erik H. Niks [3]; Nadine A. Ikelaar [3]; Bart J.A. Mertens [1]

Affiliations: [1] Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden; [2] Department of Human Genetics, Leiden University Medical Center, Leiden; [3] Department of Neurology, Leiden University Medical Center, Leiden

Introduction

In psychology, multivariate longitudinal data are frequently collected to measure a single latent construct that cannot be observed directly. Similar scenarios arise in medical research, where clinicians often rely on multiple physical assessments to evaluate patient-specific health-trajectories. Duchenne Muscular Dystrophy (DMD), a degenerative neuromuscular disorder marked by progressive loss of motor function, represents one such case. Disease monitoring typically requires repeated administration of several functional tests, each targeting different aspects of motor performance. These assessments, however, are time-consuming, physically demanding, and difficult to obtain at regular intervals. As a result, there is increasing interest in serum biomarkers that could complement or partially replace functional tests. Our objective is to identify a panel of biomarkers that jointly predict a patient's underlying disease status, modelled as a latent process representing overall functional decline, and ultimately enable inference of this latent status using biomarker data alone.

Methods

We propose a modeling framework that integrates elements from the flexible latent process approach of Proust-Lima et al. (2013) and the current-value structure of joint models1. The model links three components: (1) a longitudinal biomarker modeled through a linear mixed-effects model to estimate noise-free, subject-specific biomarker trajectories (2) a continuous latent variable characterizing each patient's underlying disease progression and capturing correlations among multiple functional assessments and (3) test-specific observation models for the physical assessments, each defined within the exponential family. The latent process includes both population-level covariate effects (e.g., age, treatment) and subject-specific deviations derived from the biomarker's random effects, enabling future prediction of disease status without requiring new functional test data. Model estimation is performed using the Integrated Nested Laplace Approximation.

Results

We illustrate the application of this modeling strategy using data from a cohort of DMD patients monitored at the LUMC.

Discussion

Although the analysis is still underway, the proposed framework has the potential to contribute to more efficient disease monitoring in DMD. By integrating biomarkers into a latent process modeling structure, the approach may help reduce reliance on frequent physical assessments, provided that the biomarker-based predictions prove sufficiently accurate.

1Rizopoulos, D. (2012). Joint Models for Longitudinal and Time-to-Event Data: With Applications in R. Chapman and Hall/CRC.

[28-P5-08] Characterizing Subgroups of Long-term Cognitive Trajectories in the Older Population

Presenter: Lydia Federmann

Authors: Lydia Federmann [1, 2]; Christian Mychajliw [3, 4]; Ulrike Sünkel [5, 4]; Didem Stark [1, 2]; Serli Kopar [1, 2]; Daniela Berg [6]; Kathrin Brockmann [7, 8]; Andreas Fallgatter [4, 5, 8]; Clint Hansen [9]; Sebastian Heinzel [9]; Walter Maetzler [9]; Florian Metzger [10, 3, 4]; Anna-Katharina von Thaler [3, 11]; Gerhard Eschweiler [3, 5, 4]; Kerstin Ritter [1, 2, 5]

Affiliations: [1] Hertie Institute for AI in Brain Health, University of Tübingen, Tübingen; [2] Tübingen AI Center, University of Tübingen, Tübingen; [3] Geriatric Centre at the University Hospital Tübingen, Tübingen; [4] Department of Psychiatry and Psychotherapy, University Hospital Tübingen, Tübingen; [5] German Center for Mental Health (DZPG), Partner Site Tübingen, Tübingen; [6] Department of Neurology, University Hospital Schleswig-Holstein and Kiel University, Tübingen; [7] Center of Neurology, University Hospital Tübingen and Hertie Institute for Clinical Brain Research, Tübingen; [8] German Center for Neurodegenerative Diseases (DZNE), University of Tübingen, Tübingen; [9] Department of Neurology, UKSH Kiel, Kiel; [10] Vitos Klinik für Psychiatrie und Psychotherapie Haina, Haina; [11] Department of Neurology, UKSH Kiel, Tübingen

Introduction

With advanced age, some individuals experience cognitive decline, while others maintain cognitive function. A better understanding of this inter-individual variability and the multidisciplinary factors influencing age-related cognitive changes may support the development of interventions that promote cognitive maintenance. Such analyses require comprehensive cognitive assessments over longtime periods and deeply phenotyped cohort data. Initiated in 2009, the TREND study recruited individuals above 48 years and has been collecting information every two years about health-related conditions, as well as cognitive, motor, and psychosocial function. Based on TREND study data, we shed light on factors associated with distinct long-term cognitive trajectories.

Methods

We clustered n=696 individuals based on features derived from their trajectories of cognitive composite scores over an average period of 13 years. Subsequently, we performed phenome-wide comparison of the resulting subgroups. Here, phenotypes spanned general physical health, motor function, mental and neuropsychological health, personality, lifestyle, diet, and genetic and biofluid markers. The comparison included the initial record of each phenotype. Multiple-testing correction was conducted across each phenotype category.

Results

In our four-cluster solution, the subgroups differ in their baseline and decline of cognitive scores. The subgroups ranged from individuals who maintain high cognitive function with advanced age, to a subgroup of individuals who showed lower baseline and faster decline of cognitive scores. The best-compared to the lowest-performing subgroup showed significantly higher gait speed, higher health-related quality of life, reported more frequent physical and cognitive stimulating activities, as well as higher intake of plant-based foods. Fewer differences were observed related to the intermediate

subgroups, yet individuals in the best-performing subgroup had significantly higher intake of plant-based foods compared to all other subgroups.

Discussion

Our clustering and phenome-wide comparison provided insights into the inter-individual variability of age-related cognitive changes. While cognitive maintenance in older adults was associated with faster gait speed and higher intake of plant-based foods, future studies may further uncover the underlying mechanisms and potential interactions across phenotypes. Together, our insights may inform future research on stratifying individuals with higher risk of early cognitive decline and guide clinical interventions to promote cognitive maintenance in the older population.

[28-P5-09] Flexible latent class mixed models reveal heterogeneous intracranial pressure trajectories in high-frequency pediatric neurocritical care data

Presenter: Francesca Graziano

Authors: Francesca Graziano [1]; Paola Rebora [2]; Stefania Galimberti [3]

Affiliations: [1] Fondazione IRCCS San Gerardo dei Tintori, 20900; [2] University of Milano-Bicocca, Monza; [3] University of Milano-Bicocca, Milan

Introduction

High-frequency longitudinal biomarkers such as intracranial pressure (ICP) in pediatric patients with traumatic brain injury (TBI) admitted to neurocritical care are characterized by irregular timing, strong within-subject dependence, and substantial heterogeneity. Conventional mixed-effects models assume population homogeneity and may fail to identify clinically relevant patterns. We applied a flexible latent class mixed modeling framework to uncover distinct ICP trajectory phenotypes.

Methods

ICP measurements were aligned to time since injury and measured at 2-hour intervals. Exploratory spaghetti plots and LOESS smoothing revealed marked non-linearity, motivating flexible modeling.

Latent class mixed models (LCMM) were applied, with time modeled via natural cubic splines (3 degrees of freedom). Model initialization was based on the single-class solution to improve convergence stability. Class-specific fixed effects enabled distinct trajectory shapes, while a parsimonious random-effects structure ensured model identifiability under high-frequency sampling. Model selection was based on likelihood criteria (AIC, BIC, SABIC), entropy, and clinical interpretability.

Results

Fifty-eight pediatric patients contributed high-frequency ICP data. The two-class model substantially improved fit over the single-class model (BIC=24104.97 vs 24914.09) and achieved excellent classification quality (entropy=0.97), identifying two well-separated trajectory groups (n=38, 65.5% vs n=20, 34.5%).

The classes showed clearly distinct temporal patterns: a predominant group with moderate ICP and gradual decline, and a second group with elevated early ICP followed by a steeper decrease. Trajectory divergence became pronounced after approximately two weeks. Baseline characteristics were broadly similar, although the higher-ICP trajectory group showed a consistent trend toward greater injury severity (Total ISS=19 vs 28.3), injury cause (road traffic incident=42% vs 15.8%, respectively in class 1 and 2) and more frequent abnormal pupillary response.

Discussion

Flexible spline-based LCMMs identified clinically meaningful ICP trajectory phenotypes not detectable with standard approaches. The use of high-frequency (2-hourly) data was key to capturing temporal heterogeneity that would likely be obscured with coarser measurements. These findings highlight how combining flexible time modeling with latent class structures can transform high-frequency monitoring data into clinically interpretable phenotypes, with direct implications for risk stratification in pediatric neurocritical care.

[28-P5-10] Predictive subgroup identification in clinical trial data using clustering of biomarker trajectories

Presenter: Johannes Vilsmeier

Authors: Johannes Vilsmeier [1]; Franz Koenig [1]; Francois Mercier [2]; Nigel Stallard [3]; Marc Vandemeulebroecke [4]; Martin Posch [1]

Affiliations: [1] Medical University of Vienna, Vienna; [2] F. Hoffmann - La Roche Ltd, Basel; [3] Warwick Clinical Trials Unit, University of Warwick, Coventry; [4] Radboud University Medical Center, Nijmegen

Introduction

Identifying subgroups of patients likely to benefit from treatment is an important goal in the exploratory analysis of randomized controlled trial data. Methods for discovering predictive subgroups are usually based on interaction terms between baseline cross-sectional biomarker measurements and treatment allocation in regression models for the clinical response, assuming that subgroup structure is reflected in cross-sectional profiles. Longitudinal biomarker trajectories measured prior to treatment initiation offer a richer source of information, as they may differ systematically across subgroups in ways cross-sectional measurements cannot capture, but incorporating them is challenging due to their high dimensionality, irregular observation, and varying numbers of time points across patients.

Methods

We simulate randomized controlled trials in which the conditional average treatment effects (CATEs) of patients depend on a latent subgroup indicator that is also predictive of the longitudinal biomarker process, modeled as a Gaussian process with subgroup-specific mean functions. Patients are divided into benefiting (CATE > 0) and non-benefiting (CATE ≤ 0) subgroups. We reduce dimensionality by clustering patients on their biomarker trajectories using latent class mixed models, functional

principal component clustering, k-means, and hierarchical clustering. Cluster indicators are embedded as treatment interaction terms in a regression model for the clinical response. Confidence and credible intervals for the CATEs are derived and an estimate of the treatment-benefiting subgroup is obtained as those patients whose lower bound on these intervals exceeds 0.

Results

Across scenarios where biomarker trajectories were informative of the underlying subgroup structure, clustering-based approaches outperformed models using only cross-sectional covariates or raw longitudinal data in terms of sensitivity, specificity, and AUC for identifying the treatment-benefiting subgroup. Notably, specificity was close to the nominal coverage of the confidence/credible intervals used for subgroup estimation. Sensitivity increased with sample size. In scenarios where subgroups were absent orbiomarker trajectories did not differ across subgroups, clustering-based and non-clustering approaches performed similarly.

Discussion

Clustering biomarker trajectories reduces the dimensionality of longitudinal measurements while preserving information on their temporal structure, and cluster membership indicators can improve identification of treatment-benefiting subgroups. This is particularly relevant in rare disease settings, where small sample sizes make dimensionality reduction essential for precise CATE estimation.

[28-P5-11] Predicting statin pharmacogenetic predispositions from clinical trajectories with knowledge graphs

Presenter: Florence Ghestem

Authors: Florence Ghestem [1]; Marc Gros La Faige [2]; Renée Le Clech [3]; Anthony Herzig [2]; Emmanuelle Génin [2, 4]; Anne-Sophie Jannot [3, 5]; Anaïs Baudot [6, 7]; Anne-Louise Leutenegger [1, 8]

Affiliations: [1] Université Paris-Saclay, UVSQ, Inserm, CESP, Villejuif; [2] Univ Brest, Inserm, EFS, UMR 1078, GGB, Brest; [3] HeKa Université Paris Cité, INRIA, INSERM, F-75015, Paris; [4] CHU Brest, Brest; [5] Banque Nationale de Données Maladies Rares (BNDMR), AP-HP, Paris; [6] Aix Marseille Univ, INSERM, MMG, Marseille; [7] CNRS, Marseille; [8] Université Paris Cité, INSERM, NeuroDiderot, Paris, UMR1141, Paris

Introduction

Pharmacogene haplotypes are associated with altered treatment efficacy or adverse response. However, pharmacogenetic predispositions to treatments remain unknown unless individuals are genotyped. Since most individuals are not genotyped before treatment delivery, clinicians rely on empirical quantitative trait measurements or adverse response declarations to adjust doses or switch treatments. Hence, we hypothesise that treatment deliveries and quantitative trait measurements recorded at multiple time points in EHR data contain information about each individual's treatment

response. Such clinical trajectories should therefore encode underlying pharmacogenetic predispositions. The present work aims to train a model that predicts pharmacogenetic predisposition from clinical trajectories.

Methods

We apply our approach to statins. We use the Constances-GOLD cohort, focusing on the 2,304 individuals with at least one lipid-modifying agent delivery (ATC3-level: C10, including statins and alternative treatments). From this dataset, we build a temporally-attributed knowledge graph to encode clinical trajectories. The knowledge graph consists of three types of nodes (individual, treatment, quantitative trait) connected by two types of edges (individual-to-treatment, individual-to-quantitative trait). Importantly, edge attributes carry temporal information and observed variable values. Our temporally-attributed knowledge graph comprises 2,304 individual nodes, 12 treatment nodes and 9 quantitative trait nodes (e.g. HDL cholesterol, diastolic and systolic blood pressure). We predict pharmacogenetic predispositions to five statins by training a GATv2 graph attention network on the clinical trajectories encoded in this knowledge graph. We derive the ground-truth binary pharmacogenetic predispositions from genotyping data. To compare our strategy with baselines, we also train tabular models (logistic regression, random forest and XGBoost). As these approaches are not temporal, they use only each individual's last recorded values.

Results

We compare the performance of the different strategies using AUC under nested cross-validation as the primary metric. We also extract attention weights of the best GATv2 model.

Discussion

Clinical trajectories carry temporal information. Therefore, we expect our temporally-attributed knowledge graph approach to outperform tabular models in predicting pharmacogenetic predispositions. Attention weight analysis may further identify the most informative clinical trajectory features for pharmacogenetic predisposition prediction. Finally, we aim to extend our comparisons by evaluating tabular models that integrate clinical trajectories.

[28-P5-12] Interpretable Temporal Modeling of Longitudinal EHRs Identifies Early Predictors of Incident Atrial Fibrillation

Presenter: Rasika Venkatesh

Authors: Rasika Venkatesh [1]; Dokyoon Kim [1]; Marylyn Ritchie [2]

Affiliations: [1] University of Pennsylvania, Philadelphia; [2] Medical University of South Carolina, Charleston

Introduction

Atrial fibrillation (AF) is the most common arrhythmia, associated with stroke and increased mortality. Early identification of high-risk patients remains a clinical priority. Longitudinal electronic

health records (EHRs) enable modeling of risk trajectories, yet most approaches lack interpretability or ignore long-term temporal structure. This study develops an interpretable deep learning (DL) framework for incident AF prediction leveraging up to 10 years of sequential patient trajectories to capture dynamic risk patterns missed by conventional cross-sectional models.

Methods

We conducted a retrospective study of 13,320 AF cases and 42,508 controls with longitudinal EHRs and genetic data in the Penn Medicine BioBank. Clinical encounters were encoded as temporal sequences in 6-month windows over a maximum 10-year period preceding AF. Comorbid conditions, including hypertension, diabetes, and heart failure, were captured with ICD-9/10 codes. Laboratory and vital measurements (cholesterol, glucose, HbA1c, eGFR, blood pressure, troponin, BMI) were aggregated into fixed-length temporal sequences. We developed baseline clinical and machine learning models including C2HEST, logistic regression (LR), and XGBoost. Longitudinal DL models (LSTM, RETAIN) were trained to capture sequential dependencies and enable attention-based interpretability. Performance was assessed using AUROC, AUPRC, balanced accuracy, F1-scores, and Brier scores. Ongoing work includes external validation in the All of US and UK Biobank datasets, incorporation of ECGs and genomic data, and further evaluation to assess clinical impact.

Results

Preliminary results indicate temporal models achieved strong performance (AUROC=0.93-0.96 AUPRC=0.89-0.92 BA=0.88-91 F1=0.81-84 Brier=0.05-0.07), outperforming C2HEST (AUROC=0.77 AUPRC=0.48), LR (AUROC=0.89 AUPRC=0.73), and XGBoost models (AUROC=0.92 AUPRC=0.84) incorporating static measurements and comorbidities. Attention analysis shows that hypertension and heart failure diagnoses, and dynamic changes in BMI, HbA1c, and systolic blood pressure, in the 12-18 month windows prior to diagnosis, contributed substantially to predictions. Sensitivity analyses across different time windows and cohort subsets supported study robustness.

Discussion

We demonstrated that interpretable temporal modeling of EHRs accurately predicts incident AF and identifies pre-diagnostic windows where risk is detectable by signals not captured in clinical scores. By leveraging multi-modal EHRs and attention-based deep learning, our framework provides insights into the dynamic patterns of comorbidities and lab measures, supporting proactive clinical intervention and facilitating integration into decision support systems for targeted screening.

[28-P5-13] Biliary Atresia-calibrated simulation of a disease progression modeling pipeline to support trial enrichment in BA

Presenter: Mehdi Gholam

Authors: Mehdi Gholam [1]; Eva Coll-de la Rubia [1]; Golnar Karimian [1]

Affiliations: [1] ProQR Therapeutics, Leiden

Introduction

Biliary atresia (BA) is a rare, rapidly progressive, life-threatening pediatric liver disease causing progressive bile duct obstruction and liver failure, with a critical unmet need for disease-modifying therapies and improved risk stratification. However, BA lacks quantitative disease progression models that bridge infancy through long-term native-liver outcomes, limiting evidence-based enrichment strategies for future disease-modifying trials. In the absence of integrated longitudinal real-world datasets, we developed a BA-calibrated simulation study to test the feasibility of the model-informed trial design pipeline.

Methods

We will generate simulated patient-level longitudinal datasets (patient, visit, clinical event, and Liver Stiffness Measurement (LSM)) informed by registry-accessible data and published BA elastography evidence. The data-generating model captures latent disease processes including cholestatic injury, fibrosis/portal hypertension, and growth and nutritional status (albumin, feeding support, etc.) as functions of time after Kasai portoenterostomy. We fit dynamic landmark and time-updated survival models plus a Bayesian multivariate hierarchical joint model linked to a major-progression endpoint of ascites, variceal bleeding, hepatopulmonary syndrome, transplant, or death. Scenario grids will pressure-test enrichment rules, event-rate assumptions, structured LSM missingness, low-quality scans, visit sparsity, null versus positive incremental LSM value, and various assumed treatment effects.

Results

This framework will generate scenario-specific estimates of model convergence, parameter recovery, calibration, discrimination, and event concentration under candidate enrichment rules. Simulations will assess whether integrating routine laboratory markers with LSM is likely to impact long term major-progression yield, identify settings in which enrichment may reduce sample-size demands, and quantify how strongly findings depend on missing LSM, scan-quality degradation, and sparse follow-up. Null-LSM scenarios will explicitly evaluate the risk of overstating biomarker utility and identify conditions under which LSM adds limited predictive value beyond standard laboratory parameters.

Discussion

The simulation-based framework will support model-informed enrichment strategy for BA trials by identifying high-risk subgroups, optimizing eligibility criteria and assessment schedules, and pre-specify sensitivity analyses prior to real-data execution. Further, it can support the quantitative justification for enrichment design, endpoint timing, and timing, facilitating more efficient and robust clinical development in rare pediatric disease.

[28-P5-14] Modelling With Informative Observations - A Simulation Study Evaluating Joint Modelling Approaches for Observational Data

Presenter: Joseph Taggart

Authors: Joseph Taggart [1]; Rhiannon K Owen [2]; James Rafferty [2]; Keith Abrams [1]

Affiliations: [1] University of Warwick, Coventry; [2] Swansea University, Swansea

Introduction

Observational data is increasingly used to evaluate the efficacy of medical interventions, but care must be taken when using such data to ensure unobserved biases do not affect conclusions. One such is the informative observation process in a trial, participants typically adhere to a strict measurement regime. With observational data, it is not possible to manage measurements this way, so participants across treatment arms may have different measurement regimes. For example, patients on different blood pressure treatments may be seen for follow-up at different rates due to reasons unrelated to the trial, such as poor health or differing co-morbidities.

Methods

One solution to this problem is joint modelling, a framework fitting longitudinal and survival models jointly to adjust for potentially informative observations and competing effects such as mortality. To evaluate the need for a joint modelling approach, a simulation study was conducted in which 100 datasets were simulated using JAGS from a joint model for blood pressure, visitation, and mortality with shared random effects. Two scenarios were investigated, one replicating the ASCOT-BPLA trial, and one with exaggerated associations between visitation and both blood pressure and treatment arm. Five models were then fitted using R-INLA/INLAjoint: a fixed-effects linear model, a linear mixed-effects model, a joint longitudinal-survival model, a joint longitudinal-visitation model, and a full joint model with both sub-models.

Results

Under the ASCOT-BPLA scenario, visitation associations were small and all five models accurately estimated the fixed-effects. Under the exaggerated scenario, the fixed-effects model produced biased estimates (empirical biases = 0.3, 0.01, 0.01 for intercept, time, and time-treatment interaction true values 5.044, -0.017, -0.001), while all four mixed-effects models estimated these accurately (biases = 0.002, 0.00001, 0.00001). The standard mixed-effects model inflated the random intercept variance, whilst the joint models decomposed the variance across sub-model components.

Discussion

An informative observation process can lead to biased fixed-effect estimates when random effects are not included. While joint models incorporating visitation and mortality sub-models correctly identify between-individual variability, this did not improve fixed-effects estimation. This suggests that a single random intercept is sufficient to account for informative visitation, without including sub-models for visitation.

[28-P5-15] Joint Bayesian modelling of antibody and PCR test data

Presenter: Larry Wang

Authors: Larry Wang [1]; Christopher Jackson [1]

Affiliations: [1] MRC Biostatistics Unit, University of Cambridge, Cambridge

Introduction

In infectious disease research, an increasing variety of data streams has become available. Modeling them jointly in a coherent and efficient manner is vital for understanding the underlying disease dynamics. In this work, motivated by the SARS-CoV-2 Immunity and Reinfection Evaluation (SIREN) study of healthcare workers in the UK, we propose a Bayesian framework that jointly models quantitative antibody and binary PCR test data. The primary aim of the model is to improve upon previous methods for estimating prevalence and incidence based solely on PCR test data. The model may also describe the evolution of immunity within individuals after exposure events.

Methods

Antibody trajectories are modeled using a smooth function that represents a boost following an exposure and a subsequent wane over time until plateau. The probability of a positive PCR test is modeled as a smooth logistic function of time since infection, and heterogeneity in immune responses is represented by covariates on antibody-related parameters. The posterior distribution was estimated using Numpyro's No U-Turn Sampler (NUTS) and Mean-Field variational inference (VI), a faster approximation. The accuracy of these algorithms is evaluated using Simulation Based Calibration (SBC) on synthetic cohorts observed over periods of 30 months.

Results

Across all simulations, NUTS yielded low bias and accurate coverage of posterior uncertainty, whereas VI suffered from systematic underestimation of posterior uncertainty. Nevertheless, NUTS proved to be computationally prohibitive for large cohort sizes, with inference taking over 2 hours for cohorts of size 1000, whilst VI took $\approx$ 2 minutes. In addition, despite under-coverage of posterior uncertainty, VI produced good point estimates, proving itself to be a useful alternative for large datasets.

Discussion

Our study demonstrates a new approach of jointly modeling antibody trajectory data and PCR test data. In ongoing work, the model is being applied to characterise incidence of SARS-CoV-2 infection among healthcare workers during 2020 - 2021, using data from the SIREN study, and determining the added value of antibody data. Further development directions include adopting a biphasic structure for the antibody trajectory to account for short term vs. long term waning, and evaluating optimal surveillance strategies for future pandemics.

[28-P5-16] Model Averaging for Individual Predictions in Joint Nonlinear Mixed-Effects Models

Presenter: Ilona Blanchard

Authors: Ilona Blanchard [1]; Emmanuelle Comets [2, 3]; Sarah Zohar [1]; Jean-Baptiste Woillard [4, 5]; Sebastien Benzekry [6, 7]; Moreno Ursino [1]

Affiliations: [1] Inserm, Inria, Université Paris Cité, UMRS 1346, Paris; [2] Univ Rennes, Inserm, EHESP, Irset (Institut de recherche en santé, environnement et travail) - UMRS 1085, Rennes; [3] Université Paris Cité and Université Sorbonne Paris Nord, Inserm, IAME, Paris; [4] Inserm P&T, U1248, Université de Limoges, Limoges; [5] CHU Limoges, Service de Pharmacologie, Toxicologie et Pharmacovigilance, Limoges; [6] COMPutational pharmacology and clinical Oncology, Centre Inria d'Université Côte d'Azur, Marseille; [7] Cancer Research Center of Marseille, Institut Paoli-Calmettes, Inserm UMR1068, CNRS UMR7258, Aix Marseille University UM105, Marseille

Introduction

Nonlinear mixed-effects models (NLMEMs) are widely used in model-informed precision dosing (MIPD), particularly in pharmacokinetic-pharmacodynamic (PK/PD) analyses where individual predictions guide treatment decisions. When multiple plausible PK/PD models coexist, structural uncertainty may substantially affect individual predictions. Existing individualised model averaging approaches have addressed either PK or PD components separately, without accounting for the hierarchical dependency inherent to joint NLMEMs.

Methods

Three individualised model averaging strategies were proposed for joint PK/PD NLMEMs, differing in how weights are propagated across hierarchical levels. Independent Individual Model Averaging (IMA Global) method, treats each PK/PD model as a single unit and assigns identical weights to both components. Independent Individual Model Averaging (IMAInd) method, weights PK and PD components separately, allowing model contributions to vary at each level. Conditional Individual Model Averaging (IMACond) method, uses the averaged PK profile as a fixed input for all candidate PD models. This strategy required re-estimation of PD parameters while maintaining population parameters and updating individual estimates. Individual model weights were primarily based on the individual likelihood. To reduce overfitting in sparse-data settings, we introduced an uncertainty-penalised weighting scheme (SIC-based) incorporating a term proportional to the log-determinant of the posterior variance-covariance matrix of the individual parameter estimates.

Performance was evaluated through a tacrolimus simulation study combining three published PK models with three PD models across three scenarios: S1, each patient is generated by a single joint model S2, PK and PD are selected independently for each patient S3 uses mixture-generated PK and PD. Four observations per subject were used for estimation. Predictive accuracy was evaluated using relative bias (rBias) on 120 concentration time points.

Results

Under S1, all methods showed negligible median rBias (~0%). Under S3, IMAGlobal exhibited a +7% bias, whereas IMACond remained centered (median rBias -1.1%). IMAInd reduced bias propagation but showed increased variability. SIC-based weights improved robustness in complex scenarios, particularly under sparse sampling.

Discussion

The main contribution of this work is the comparative evaluation of hierarchical individual model averaging strategies under structural misspecification. Accounting for PK/PD dependency improves predictive stability, with IMACond providing promising performances. The framework extends to general hierarchical NLMEM settings, beyond PK/PD applications.

[28-P5-17] Efficient Bayesian inference for non-linear association structures in joint models: A hierarchical approach via INLA

Presenter: Denis Rustand

Authors: Denis Rustand [1]

Affiliations: [1] Inserm U1219, University of Bordeaux, Bordeaux

Introduction

Joint models characterizing the association between biomarker trajectories and clinical events typically assume a linear relationship. This restrictive assumption frequently fails to capture complex biological mechanisms, such as U-shaped risk profiles. Existing flexible extensions that rely on Markov Chain Monte Carlo (MCMC) sampling suffer from significant computational burden and numerical instability, preventing routine adoption.

Methods

We propose a fast and stable hierarchical framework to estimate non-linear association structures using Integrated Nested Laplace Approximations (INLA), implemented in the open-source INLAjoint R package. Our approach decomposes the scaling effect of the longitudinal marker into a parametric baseline (constant and linear components) and a flexible, data-driven smooth deviation. This deviation is modeled via an orthogonal basis derived from a second-order random walk. This nested hierarchy allows users to incrementally adapt model flexibility and verify the linearity assumption using standard information criteria.

Results

Simulation studies demonstrated that our method accurately recovers complex non-linear trajectories while avoiding overfitting when the true association is simpler. We applied the framework to 42,403 individuals from the Health and Retirement Study to simultaneously assess the joint association of body mass index (BMI) current value and current slope with all-cause mortality. The analysis revealed a U-shaped risk for BMI value and a distinct non-linear S-shaped effect for the rate of weight change, indicating an increased mortality risk for weight loss and a protective effect for moderate weight gain. Model estimation for this large cohort required less than 30 minutes.

Discussion

The proposed INLA-based framework successfully overcomes the computational bottlenecks and convergence challenges of sampling-based estimation. By scaling efficiently to large datasets and allowing the simultaneous estimation of multiple non-linear shared components, it provides a practical, accessible tool for biostatisticians to routinely diagnose and model complex dynamic risk profiles.

[28-P5-18] Bayesian Spatial Joint Latent Class Modeling for Longitudinal and Survival Data with Dynamic Risk Prediction

Presenter: Zahra Seyedi Lahroudi

Authors: Taban Baghfalaki [1]; Mojtaba Ganjali [2]; Zahra Seyedi Lahroudi [1]; Asri Maharani [1]

Affiliations: [1] The University of Manchester, Manchester; [2] Shahid Beheshti University, Tehran

Introduction

Joint Latent Class Models (JLCMs) integrate latent class analysis with joint modeling of longitudinal and survival data to capture unobserved heterogeneity. They are widely applied in epidemiological studies to identify meaningful subgroups and enable personalized risk assessment. The English Longitudinal Study of Ageing (ELSA) provides extensive longitudinal data on biomarkers, health outcomes, and social factors in older adults, enabling comprehensive evaluation of aging trajectories across diverse populations.

Methods

We propose a Bayesian spatial JLCM that jointly models longitudinal biomarker trajectories and time-to-event outcomes while accounting for latent subgroups and regional variation. Individuals are grouped by geographic regions, and spatial correlation is incorporated via ICAR priors on regional random effects, which influence both latent class probabilities and survival hazards. Individual-level covariates included in the model are those shown to be significant in previous ELSA studies, including age, sex, education, socioeconomic status, lifestyle factors (smoking, physical activity), and health conditions such as cardiovascular and metabolic disorders. Parameters are estimated using Hamiltonian Monte Carlo (HMC) to efficiently sample from the joint posterior distribution. The model supports dynamic risk prediction, updating individual risk profiles as new longitudinal measurements are observed over time, providing clinically relevant, personalized insights. Predictive performance is evaluated using time-dependent AUC and Brier Score, offering quantitative assessment of discrimination and calibration. Additionally, simulation studies are conducted to assess the model's ability to recover latent classes, spatial effects, and predictive accuracy under varying cohort sizes, censoring patterns, and measurement schedules.

Results

Simulation studies confirm that the Bayesian spatial JLCM reliably recovers latent structure, spatial effects, and provides well-calibrated predictions across diverse scenarios. Application to ELSA

data identifies latent subgroups with distinct biomarker trajectories and survival risks while capturing regional heterogeneity. Dynamic risk prediction demonstrates early identification of high-risk individuals and highlights regions with elevated risk. Covariate effects align with known significant predictors, including age, socioeconomic status, and health behaviors.

Discussion

These findings allow for both personalized risk profiling and potential targeting of high-risk areas. Overall, the model provides actionable insights, supports targeted interventions, and demonstrates strong predictive accuracy, interpretability, and practical relevance for aging research.

[28-P5-19] Landmarking with Latent Class Mixed Models for Dynamic Prediction of Time-to-event Data with Heterogeneous Biomarker Trajectories

Presenter: Victor Velasco-Pardo

Authors: Victor Velasco-Pardo [1]; Nathan Constantine-Cooke [1]; Charlie W. Lees [1, 2]; Catalina Vallejos [1, 3]

Affiliations: [1] Institute of Genetics and Cancer, University of Edinburgh, Edinburgh; [2] Edinburgh IBD Unit, Western General Hospital, Edinburgh; [3] Health Data Research UK, London

Introduction

The increased availability of electronic health records (EHR) has opened new possibilities for monitoring health trajectories across large and diverse patient populations over time. Longitudinal measurements of biomarkers enable dynamic risk prediction of time-to-event outcomes. Most existing dynamic prediction methods assume that longitudinal trajectories are homogeneous across patients. This assumption may not hold in practice if unobserved confounders give rise to latent subgroup structure and heterogeneous trajectories, particularly in real-world EHR data. The only heterogeneity-aware method is the joint latent class model (JLCM), which is computationally too demanding for the large-scale datasets typical of EHR settings. The landmarking framework offers a flexible, modular approach to dynamic prediction, yet no existing landmarking method accommodates latent heterogeneity in biomarker trajectories.

Methods

We introduce a heterogeneity-aware extension to landmarking that leverages the latent class mixed model (LCMM) to summarise biomarker trajectories. The approach is made available through landmaRk, a modular R package designed to give users flexibility in specifying each component of the landmarking analysis. Discrimination and calibration performance are assessed using time-dependent AUC and Brier score, respectively, and benchmarked against standard landmarking approaches and JLCMs.

Results

In simulation studies, we demonstrate that the proposed method improves discrimination and calibration compared to existing landmarking approaches in scenarios with latent subpopulation structure and achieves performance competitive with JLCM while requiring considerably less computation.

Discussion

Our method enables dynamic prediction that explicitly acknowledges latent trajectory heterogeneity, while remaining computationally efficient for large-scale datasets. Its implementation in the landmaRk package can easily accommodate future methodological extensions.

P6 - Evidence and Decision Sciences

Poster area: P6

[28-P6-01] Developing methods guidance for health researchers: literature review of proposed and applied strategies

Presenter: Malena Chiaborelli

Authors: Malena Chiaborelli [1]; Julian Hirt [2, 3, 4, 5]; Matthias Briel [1]; Frances Shiely [6, 7, 8]; Anne-Laure Boulesteix [9, 10]; Gordon Guyatt [11]; Frank Bretz [12, 13]; David Moher [14, 15, 16]; Derek Stewart [17]; Stefan Schandelmaier [1, 18]

Affiliations: [1] Department of Clinical Research, CLEAR-Methods Centre, University Hospital Basel and University of Basel, Basel; [2] Department of Clinical Research, Faculty of Medicine, University of Bern, Bern; [3] Department of Health, Eastern Switzerland University of Applied Sciences, St Gallen; [4] Institute of Health, Midwifery and Nursing Science, Medical Faculty, Martin Luther University Halle-Wittenberg, Halle; [5] JBI Centre for Evidence-Based Healthcare Eastern Switzerland, University of Adelaide, Adelaide; [6] TRAMS (Trials Research and Methodologies Unit), HRB Clinical Research Facility, University College Cork, Cork; [7] HRB Trials Methodology Research Network (TMRN), University College Cork, Cork; [8] School of Public Health, University College Cork, Cork; [9] Institute for Medical Information Processing, Biometry and Epidemiology, Faculty of Medicine, LMU Munich, Munich; [10] Munich Center for Machine Learning, Munich; [11] Departments of Medicine and Health Research Methods, Evidence, and Impact, McMaster University, Hamilton; [12] Novartis Pharma AG, Basel; [13] Medical University of Vienna, Center for Medical Data Science, Institute of Medical Statistics., Vienna; [14] Methodological and Implementation Research Program Centre for Journalology, Ottawa Hospital Research Institute, Ottawa; [15] School of Epidemiology and Public Health, Faculty of Medicine University of Ottawa, Ottawa; [16] Institute of Health Policy, Management & Evaluation, Dalla Lana School of Public Health, University of Toronto, Toronto; [17] Health Research Board - Trials Methodology Research Network, University of Galway, Galway; [18] MTA-PTE Lendület 'Momentum' Evidence in Medicine Research Group, Medical School, University of Pécs, Pécs

Introduction: Health researchers benefit from methods guidance (MG) for planning, conducting, analyzing and interpreting research studies. How to develop MG that is trustworthy and researcher-friendly is unclear existing strategies for developing MG are fragmented and lack a coherent framework. The STrategies for developing REseArch Methods guidance (STREAM) project aims to provide such a framework for developers of MG.1 As a first step, we reviewed the literature to identify proposed and applied strategies for developing MG and explored possible knowledge gaps. Methods: We conducted two literature reviews. First, a scoping review of the methodological literature to identify proposed strategies for developing MG. Second, a review of a purposeful sample of recently published MG articles to examine strategies applied in practice. We applied a qualitative content analysis to both sets of publications using one shared codebook. In addition, to identify possible knowledge gaps, we contrasted the findings with the established development processes for clinical practice guidelines and reporting guidelines. Results: The first review included 17 publications and the second review 102 MG articles. We identified possible development steps including assessing the need for MG, establishing a working group, considering the findings from methodological studies, gathering input from interest-holders, testing the new MG, and disseminating and implementing MG. In addition, we identified various approaches to perform the steps including systematic reviews of previous guidance, reviews of methodological practice, simulation studies, surveys or interviews of interest-holders, consensus methods, and public review. Mapping these findings with established development processes identified areas that have so far received limited attention within MG development, including the concept of methodological evidence, judging the quality of methodological evidence, reporting of MG, and approaches to implementation. Discussion: This study provides an overview of possible strategies for developing MG, highlights needs for conceptual development, and provides a structured starting point for guidance for developing MG. Reference: 1. Chiaborelli,

M. et al. STrategies for developing REseArch Methods guidance (STREAM): protocol. J. Clin. Epidemiol. 112184 (2026).

[28-P6-02] Misquotation errors of articles critical of using the Newcastle-Ottawa Scale in systematic reviews. The disaster continues.

Presenter: Andreas Stang

Authors: Andreas Stang [1]; Hendrik Damm [2]; Henning Schäfer [3]; Christoph M. Friedrich [4]; John D. Murphy [5]; Charles Poole [6]

Affiliations: [1] Institut für Medizinische Informatik, Biometrie und Epidemiologie, Universitätsklinikum Essen, Essen; [2] Department of Computer Science, University of Applied Sciences and Arts Dortmund (FHDO), Dortmund; [3] Institute for Transfusion Medicine, University Hospital Essen, Essen, Germany, Essen; [4] Fachhochschule Dortmund Fachbereich Informatik, Dortmund; [5] Data Science Team, Odysseus Data Services, Washington DC; [6] Department of Epidemiology, University of North Carolina, Chapel Hill

Introduction: In 2010, Stang criticized the Newcastle-Ottawa scale (NOS), which is used to assess the quality of observational studies, for various reasons and advised against its use in systematic reviews. In 2018, Stang et al. (2018) found that an estimated 98% of publications had cited Stang (2010) in support of using the NOS for that purpose. The aim was to update that citation analysis and begin a similar analysis of publications citing Stang et al. (2018). Material and Methods: We used Scopus to search for publications citing Stang 2010 and reviewed a random sample after the appearance of Stang et al. (2018) and also reviewed all publications citing Stang (2018). Of the references citing Stang (2010), open access full texts of references were downloaded and analyzed by use of large language models to automatically categorize quotation errors. Results: From 2010 through 2024, we found 12,803 citations of Stang (2010). The annual number of publications of systematic reviews of observational studies and of Stang (2010) citations rose steadily from 2010 to 2024. The estimated proportion of systematic reviews that cited Stang 2010 rose steadily from 2010 to 2020, then remained constant until 2023, only to rise again until 2024, when it reached its maximum thus far of 28%. Among the available 5,959 full texts of 12,803 publications citing Stang (2010), the estimated percentage that correctly quoted Stang 2010 was 1.6%. Overall, 86% (95%CI: 79-92) of publications after 2018 quoted Stang (2010) in favor of using the NOS in a systematic review. The prevalence of major quotation errors of Stang et al. (2018) in systematic reviews was 80% (95%CI: 66-91). Both publications (2010 and 2018) were also incorrectly quoted for specifying specific categorizations of NOS scores. Discussion: Even after the second critical publication 2018 on using the NOS in systematic reviews, major quotation errors remain very common. It remains the responsibility of authors to carefully read cited references. In the future, artificial intelligence approaches may help to automatically detect citation and quotation errors.

[28-P6-03] Statistical idealism: Registered Reports for clinical trials

Presenter: Judith ter Schure

Authors: Judith ter Schure [1]

Affiliations: [1] Amsterdam UMC, Amsterdam

Introduction

Despite all improvements since Doug Altman's 'The scandal of poor medical research', it is still not hard to find a 'high-impact' medical journal paper that has serious statistical flaws. We must not accept this situation, not just complain, and look into new solutions. The Registered Report (RR) is a publishing format that focuses on pre-data review and in-principle acceptance - promising to prevent poor rationale, inappropriate methods, selective reporting and publication bias.

Methods

The RR format and implementations with joint funder/ethics review can help us rethink our roles: (1) Assess the effectiveness of statisticians' duplicated reviews (funder, ethics, journal), specifically for investigator-initiated clinical trials, and how we get from 'firefighting' to 'fireproofing'. (2) Recognize the option to take statistical review out of the hands of medical journals, when our involvement signals quality. Can we build an amber-proof place where medical research wants to be in? (3) Examine platforms outside of the journal system, like Peer Community In Registered Reports (PCI-RR). (4) Identify opportunities for micro-activism in our roles as clinical trial collaborators, funding reviewers/committee members, ethical board members, (prospective) DSMB members, and journal editors and reviewers. (5) Consider organizing, as a working group within the ISCB or by focusing on individual action that is actively shared.

Results

These five points up are up for discussion, and will be illustrated by a case study of a clinical trial that went through open Stage-1 peer review at PCI-RR simultaneously with ethics review, all scheduled and completed within weeks. Most interesting was the reliable timeline for complete review and the optionality of submissions to medical journals. To reflect on organizing around PCI-RR, I will also summarize advocacy by the Clinical Research working group of the European Association for Clinical Pharmacology and Therapeutics.

Discussion

We might want to stop playing in the backyard of the medical prestige economy. PCI-RR could offer what we need: our own backyard to invite our clinical colleagues, starting with clinical trials.

Altman, D. G. (1994). The scandal of poor medical research. Bmj, 308(6924), 283-284.

Collins, G. S., & Deeks, J. J. (2025). The making of a statistician: Doug Altman. Bmj, 391.

[28-P6-04] Academic Integrity in the Age of Biomedical AI: Analysis of Retracted Articles from the Perspective of Scoping Review and E-PICOS Reviewer

Presenter: EMİNE Kanik, Burak Akicier

Authors: EMİNE Kanik [1]; Burak Akicier [1]

Affiliations: [1] AB Health Tech MedicReS, Houston

Introduction

The use of artificial intelligence and Large Language Models in biomedical research is revolutionizing data analysis and reporting processes while simultaneously creating new challenges for academic integrity and methodological rigor. The increasing number of retracted articles indicates a need for stronger oversight mechanisms to preserve the reliability of scientific literature during this transition period.

Methods

In this study, a comprehensive literature search was conducted to identify retracted articles, and 131

were analyzed using a Scoping Review methodology. During the analysis, the 'Statistical Peer Review Checklist' was utilized to evaluate methodological and statistical errors systematically. Furthermore, the potential role of the E-PICOS Reviewer platform in detecting these errors pre-publication and its compliance with PRISMA standards were evaluated.

Results

It was found that 35.9% (n=47) of the analyzed articles were retracted directly due to statistical or methodological reasons. Particularly in the post-2023 period, a dramatic increase in retractions due to undisclosed AI use and associated methodological inconsistencies was observed. The most affected areas were identified as study design (55%), reporting of results (34%), and numerical data verification (34%).

Discussion

The findings emphasize the critical importance of structured checklists and automated oversight tools (such as E-PICOS Reviewer) for maintaining methodological transparency and reproducibility in the age of biomedical AI. Successful management of this transition period in scientific publishing is only possible by supporting technological innovations (AI) with technological oversight mechanisms.

[28-P6-05] Comparing Existing Methods to Assess the Multiplicity of Analysis Strategies: A Pharmacoepidemiological Case Study Using RMST-Based Comparisons of Multiple Treatments

Presenter: Moritz Pamminger

Authors: Moritz Pamminger [1]; Sinan Necdet Cevirme [1]; Michael Kammer [1]; Susanne Strohmaier [2]; Daniela Dunkler [1]

Affiliations: [1] Institute of Clinical Biometrics, Center for Medical Data Science, Medical University of Vienna, Vienna; [2] Department of Epidemiology, Center for Public Health, Medical University of Vienna, Vienna

Introduction

When addressing a particular research question using observational data, many decisions must be made during the conceptualization and execution of the statistical analysis plan. This multiplicity of analysis strategies is a well-known problem that leads to high variation in research findings and associated low reproducibility, since each decision can lead to different results, even if each decision on its own was scientifically justifiable.

Methods

Several approaches for reporting and visualising variation in effect estimates have been proposed. The vibrations of effects framework by Patel (2015) focuses on how effect estimates change across different model specifications and provides intuitive visualisations. Specification curves by Simonsohn et al. (2020) offer an alternative approach by systematically displaying results across a range of analytical decisions, including variable selection and model choices.

We exemplify these methods in a pharmacoepidemiological case study, comparing the effect of three different P2Y12 inhibitors (clopidogrel, ticagrelor, prasugrel) regarding time to first recurrence of acute coronary syndrome (ACS) or death in a cohort of 12,000 stented ACS patients. We obtained data from the Austrian Health Insurance Fund, including basic demographic information, information on all hospitalisations between 2019 and 2023 and all drug redemptions in that time period.

We computed 3-year restricted mean survival times (RMSTs) and 3-year survival probabilities across various justifiable specifications and visualised them using the vibration of effects and specification curve framework.

Results

The resulting RMSTs for the main specification were 938 (95%-CI: 922-953) for clopidogrel, 959 (942-979) for prasugrel and 955 (944-966) for ticagrelor, while other specifications showed considerable variation.

Discussion

Specification curves are typically restricted to displaying a single estimate per specification (e.g. a regression coefficient with its confidence interval), whereas our setting requires the simultaneous consideration of survival probabilities or RMSTs across multiple treatments. Similarly, the vibration of effects framework has traditionally focussed on investigating variation in hazard ratios and corresponding p-values due to differences in adjustment sets. We extend the approaches beyond their typical scope by addressing challenges that arise when investigating alternative outcome measures for time-to-event data in settings where multiple contrasts are of interest. This project is funded by ÖAW DATA 2023-32.

[28-P6-06] Reverse-Bayes Methods for Assessing the Credibility of Risk Estimates in Cardiovascular Research

Presenter: Keiko Mori

Authors: Keiko Mori [1]; Ayano Takeuchi [2]

Affiliations: [1] Chuo University, Graduate School of Science and Engineering, Saitama; [2] Chuo University, Graduate School of Science and Engineering, Tokyo

Introduction

Null hypothesis significance testing (NHST) remains the dominant framework in biomedical research. However, it has been widely criticized for its reliance on dichotomous interpretation (significant vs. non-significant), sensitivity to sample size, and instability of p-values across repeated studies. These limitations hinder the assessment of the reliability and reproducibility of research findings. Bayesian approaches have been proposed as alternatives, as they allow uncertainty to be quantified through posterior distributions that incorporate prior information. However, Bayesian methods face challenges related to the subjective specification of prior distributions, known as the "problem of priors." Reverse-Bayes methods have recently gained attention as a promising framework that addresses this issue by inverting the Bayesian paradigm: instead of specifying priors, they determine the required prior plausibility (e.g., the sceptical prior) necessary to support or challenge observed findings. While these methods have been increasingly discussed in methodological research, their application in applied biomedical and clinical research is still emerging.

This study aims to evaluate the application of reverse-Bayes methods for the quantitative assessment of the credibility of risk estimates in cardiovascular disease research.

Methods

We propose a reverse-Bayes framework based on the Analysis of Credibility approach. Utilizing summary-level statistics (effect estimates and their uncertainty)), we derive the minimum prior strength-specifically the sceptical prior required th conflict with statistically significant findings or support non-significant findings. These quantities are interpreted in terms of the plausibility of prior assumptions and evaluated in relation to existing clinical knowledge.

Cardiovascular disease provides a particularly relevant context for this approach, as risk assessment frequently relies on biomarkers with modest effect sizes. In such settings, statistical significance is notoriously sensitive to sample size and study design, leading to variability in p-values and inconsistencies across studies. Furthermore, the large number of potential biomarkers increases the risk of spurious findings.

Results

Reverse-Bayes methods offer a quantitative framework to assess the credibility of statistical findings by explicitly incorporating prior plausibility.

Discussion

This approach has the potential to complement NHST and improve the interpretation and robustness of evidence in cardiovascular research.

[28-P6-07] Evaluating Large Language Models for Title and Abstract Screening in Reviews: A Study Within A Review

Presenter: Stella Erdmann

Authors: Stella Erdmann [1]; Arvid Holsmoelle [2]; Meinhard Kieser [2]

Affiliations: [1] Institute of Medical Biometry, University of Heidelberg, Heidelberg, Heidelberg; [2] Institute of Medical Biometry, University of Heidelberg, Heidelberg

Introduction

Large language models (LLMs) have demonstrated substantial potential to support and automate key steps of systematic reviews.[1] With the release of the generalized pretrained transformer (GPT-)5 of openAI, initial findings indicate improved reasoning compared to earlier models. Static models, e.g., from the Llama family of Meta on the other hand offer advantages in reproducibility, transparency, governance, and local data control. This study aims to systematically evaluate the feasibility, accuracy and efficiency of static and dynamic LLMs when embedded in a review workflow.

Methods

A Study Within A Review (SWAR registered in the SWAR Store: SWAR ID 66)[2] is embedded in an ongoing scoping review on causal inference methods using real-world data in healthcare research. Review tasks - including title and abstract screening and structured data extraction - are conducted in parallel using standard dual-reviewer procedures as gold standard and LLMs: Title and abstract screening is performed by GPT-5 and GPT-5-mini (via API), LLAMA 3.1 8B Instruct and Llama 3.3 70B. The screening decision (include/exclude/unclear) is based on four screening questions, all of which must be answered 'yes' for a paper to be included. Sensitivity, specificity, agreement rates and time expenditure are compared between approaches.

Results

The title and abstract screening is conducted on 2,226 deduplicated records identified through database searches. Preliminary findings indicate that LLM screening substantially reduces workload, while maintaining high sensitivity, specificity and agreement with human inclusion decisions. However, occasional misalignment was observed whereby the LLM correctly answered all individual screening questions but still produced an incorrect overall screening decision.

Discussion

This SWAR provides empirical evidence on the practical integration of LLMs into systematic review workflows, while also contributing to the emerging evidence base on best practices in prompt engineering. By evaluating both efficiency gains and methodological risks, this study contributes to the responsible adoption of LLMs in review workflows.

References

1. Guo, E., et al., Automated paper screening for clinical reviews using large language models: data analysis study. Journal of Medical Internet Research, 2024. 26: p. e48996.

2. Devane, D., et al., Study within a review (SWAR). Journal of Evidence-Based Medicine, 2022. 15(4): p. 328.

[28-P6-08] Artificial Intelligence for Interview Coding: A Four-Study Evidence Review

Presenter: Gabriel Hess

Authors: Felipe Ricci [1]; Kerstin Denecke [2]

Affiliations: [1] UNIFESP, São Paulo; [2] Berner Fachhochschule, Biel

Introduction

Early-stage health innovation discovery relies on interviews and qualitative synthesis, yet these steps are time-intensive, difficult to standardize, and vulnerable to over-interpretation. Large language models (LLMs) may accelerate transcript-to-insight workflows (summaries, candidate codes, draft themes), but can also introduce hallucinations, unreliable quotation, and cross-cultural misinterpretation-risks that may distort problem framing and prioritization.

Methods

We conducted an exploratory comparative multiple-case synthesis of four peer-reviewed, head-to-head evaluations comparing LLM-supported qualitative analysis with human qualitative analysis or expert coding. Studies were identified through a targeted scan and included if they (i) reported direct LLM-versus-human comparisons, (ii) addressed discovery-relevant tasks (thematic analysis/summarization, inductive coding, or quote selection), and (iii) provided sufficient methodological detail to extract performance signals, failure modes, and safeguards. We extracted task scope, data context/language, evaluation approach and metrics, reported benefits, observed failures (quote fidelity, grounding, cultural drift), and recommended or implied safeguards, then synthesized cross-case patterns.

Results

LLMs performed best on descriptive structuring (theme listing, clustering, high-level summaries) and weakest on interpretive work requiring implicit meaning and sociocultural nuance. One study reported highly variable performance in selecting quotations that strongly supported assigned sentiments, raising auditability concerns when "quotes-as-evidence" underpin conclusions. Another study showed high overall agreement under a structured rubric, yet highlighted risks of coherent elaborations not clearly traceable to source text. A Japanese clinical interview study reported high agreement for descriptive themes but markedly lower agreement for culturally embedded interpretation. An inductive-coding benchmark found complementary human/LLM error profiles, suggesting hybrid workflows may help but require explicit governance.

Discussion

LLMs are best positioned to accelerate mechanical steps in discovery, not to replace human interpretation. Minimum safeguards include (i) traceability from themes to transcript spans, (ii) extractive-only quoting with human verification, (iii) explicit human ownership of interpretive synthesis, and (iv) versioned prompts and audit logs for reproducibility. References: Mehta SD, Paul S, Awiti E, et al. Evaluation of large language models within GenAI in qualitative research. Scientific Reports 202515:34993. doi:10.1038/s41598-025-18969-w. Tai RH, Bentley LR, Xia X, et al. An examination of the use of large language models to aid analysis of textual data. International Journal of Qualitative Methods 202423:16094069241231168. doi:10.1177/16094069241231168.

The full paper pertaining to this abstract will be published in the GMDS journal "Medizinische Informatik, Biometrie und Epidemiologie (MIBE)."

[28-P6-09] An Integrated AI-Assisted Pipeline for Translating Treatment Effects Across Correlated Endpoints Using Automated Clinical Trial Data Extraction

Presenter: Nusrat Rabbee, Nusrat Rabbee

Authors: Nusrat Rabbee [1]; Wenjun Zheng [2]; Bingying Dai [2]

Affiliations: [1] Regeneron Pharmaceuticals, Inc, Clifton; [2] Regeneron Pharmaceuticals, Inc, tarrytown

Introduction

Translating treatment effects across correlated endpoints in meta-analysis requires aggregating data from multiple clinical trials. Manual extraction from ClinicalTrials.gov is labor intensive, error prone, and limits the scalability of such analyses across therapeutic areas.

Methods

We present an integrated two-stage pipeline. First an AI tool automatically extracts and structures trial-level data from ClinicalTrials.gov across a target therapeutic area. Second we apply SLIM - Stacked Linear Mixed Effects Models - to the aggregated correlated endpoints, providing a computationally efficient alternative to Bayesian MCMC approaches while preserving observed treatment effect relationships across endpoints.

Results

We demonstrate the pipeline applied to a therapeutic area of interest, illustrating how automated extraction combined with SLIM enables scalable and reproducible translation of treatment effects across correlated endpoints.

Discussion

his integrated pipeline advances surrogate endpoint research by combining AI-assisted clinical trial data collection with a statistically rigorous and accessible analytical framework, with broad applicability across therapeutic areas.

[28-P6-10] IQWiG meta-analyses database: A presentation of the database and overview of two application examples

Presenter: Leonie Schürmeyer

Authors: Leonie Schürmeyer [1]; Sibylle Sturtz [1]; Samuel Pawel [2]; Saverio Fontana [2]; Frank Weber [1]; Guido Skipka [1]; Leonhard Held [2]; Ralf Bender [1]

Affiliations: [1] Institute for Quality and Efficiency in Health Care (IQWiG), Cologne; [2] University of Zurich, Zurich

Introduction

The Institute for Quality and Efficiency in Health Care (IQWiG) is an independent institute examining the benefits and harms of medical interventions for patients in form of scientific reports. In doing so, IQWiG often employs meta-analyses and maintains a database containing the study-specific data of all meta-analyses from its published reports since the institute was founded in 2004.

Methods

We present the database using descriptive analyses and information on various aspects such as the number of included studies or the effect measure used. Furthermore, the process of creating and maintaining the database as well as application examples are described.

Results

So far, the database has been used to derive appropriate prior distributions for the heterogeneity parameter in Bayesian meta-analyses at IQWiG [1]. The large empirical database can also be useful for addressing various other research questions. Based on existing research on combined p-value functions [2], we are investigating whether methods based on Edgington's p-value combination method are more suitable than the currently employed methods for meta-analyses of two studies based on the common-effect assumption. Methods based on Edgington are calibrated under between-trial homogeneity but adapt the width of their confidence intervals to the observed discrepancy between trial results.

For both examples, we present key results based on the empirical IQWiG database.

Discussion

The database provides a vast repository of data already available in the health technology assessment sector, thereby offering a valuable foundation for conducting scientific analyses and investigating meta-analytic methods based on this empirical data. It offers a basis for realistic empirical analyses and provides a reasonable foundation for simulation studies. We plan to maintain the database and publish it in the future.

[1] Lilienthal, J., Sturtz, S., Schürmann, C., Maiworm, M., Röver, C., Friede, T., & Bender, R. (2024): Bayesian random-effects meta-analysis with empirical heterogeneity priors for application in health technology assessment with very few studies. Research Synthesis Methods, 15(2), 275-287.

[2] Held, L., Hofmann, F., & Pawel, S. (2025): A comparison of combined p-value functions for meta-analysis. Research Synthesis Methods, 16(5), 758-785.

[28-P6-11] LSRUpdateR: Making informed decisions on when to update a living systematic review

Presenter: Clareece Nevill

Authors: Clareece Nevill [1]; Nicola Lindson [2]; Alex J Sutton [1]; Ellesha Smith [1]; Jamie Hartmann-Boyce [3]; Suzanne Freeman [1]; Nicola Cooper [1]

Affiliations: [1] Biostatistics Research Group, Division of Public Health and Epidemiology, University of Leicester, Leicester; [2] Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford; [3] Department of Health Policy and Promotion, University of Massachusetts Amherst, Amherst

Introduction

Living systematic reviews (LSRs) keep evidence up to date by regularly searching and incorporating new studies. A 2023 systematic methodological survey of LSRs by Nevill et. al. highlighted that whilst LSRs are increasing, approaches regarding when to publish updates vary considerably and need further development. Whilst some reviews updated at set frequencies, others used 'triggers' -updating when new studies change the conclusions or certainty of evidence. However, because it is often impossible to know whether conclusions or certainty of evidence has changed without first conducting the update, decisions are often based on intuition. This project aimed to develop a tool that provides transparent information to support and justify such decisions.

Methods

The 2023 methodology review indicated that update triggers commonly focus on (1) overall conclusions, assessed through pooled effect estimates, and (2) certainty of evidence via GRADE. These domains formed the basis of an update tool, for which associated quantitative measures, such as confidence intervals and overlap metrics, were investigated and developed. The tool's performance was assessed using the UK-based Cochrane LSR on e-cigarettes for smoking cessation (https://www.cebm.ox.ac.uk/research/electronic-cigarettes-for-smoking-cessation-cochrane-living-systematic-review-1). It was applied retrospectively to examine how well it predicted changes in conclusions or certainty of evidence compared with the real published updates.

Results

When applied to the e-cigarette LSR data, spanning 2020 - 2025, the tool suggested twelve updates based on predicted changes to primary outcomes, whereas five were published in practice. The tool successfully identified the reported changes in primary outcomes regarding conclusions and certainty of evidence, but appeared to be more sensitive and granular regarding when to conduct updates. An R function, LSRUpdateR, was developed that recommends to users whether a review should be updated, giving details of what changes would likely occur from including newly found studies.

Discussion

LSRUpdateR was able to make similar update decisions to review teams, whilst offering automation and transparent decision rules. Whilst LSRUpdateR needs further refinement, in terms of additional features as well as how to use in practice, its use could improve reproducibility and resource allocation. LSRUpdateR encourages LSR teams of any experience to make informed update decisions robustly and consistently.

[28-P6-12] A new PCO pyramid to reflect an extended PICO framework for network meta-analysis

Presenter: Fani Apostolidou-Kiouti

Authors: Fani Apostolidou-Kiouti [1]; Anna-Bettina Haidich [1]; Dimitrios Mavridis [2]; Adriani Nikolakopoulou [1]

Affiliations: [1] Department of Hygiene, Social-Preventive Medicine and Medical Statistics, School of Medicine, Aristotle University of Thessaloniki, Thessaloniki; [2] Department of Primary Education, University of Ioannina, Ioannina, Greece, Ioannina

Introduction

The PICO framework has been the mainstay for research question formulation in evidence synthesis, delineating clear definitions for population, intervention, comparator, and outcomes. However, clinical questions have become increasingly complex, involving multiple treatments, varying doses, effect modifiers, and multiple outcomes. Methodological advances in meta-analysis have emerged to address these challenges, extending traditional PICO components. In this study, we review these advances and assess their uptake in published network meta-analyses (NMAs), with the aim of informing an extended conceptual framework.

Methods

We conducted a comprehensive search for published NMAs between January 2020 and November 2025 using database searches and citation tracking. Studies were assessed for the use of advanced NMA models (component NMA, dose-response NMA, model-based NMA, population adjustment methods in NMA, multivariate NMA models). Study and methods characteristics were extracted and summarized descriptively.

Results

A total of 141 studies were identified. Frequentist component NMA was the most commonly used method (n = 56, 39.7%), followed by Bayesian dose-response NMA (n = 31, 21.9%). The least applied model was multivariate NMA (n = 3, 2.1%). Most of the identified studies addressed mental health and neurology (n = 52, 36.8%).

Discussion

There is growing use of advanced NMA methodologies to address increasingly complex clinical research questions. Based on these findings, we propose a conceptual extension of the traditional PICO framework in the form of a PCO pyramid. This framework allows each component to expand from a single element to multiple elements (1 to n), accommodating multiple populations, interventions, and outcomes. As complexity increases, new sources of variability arise, requiring renewed considerations of fundamental meta-analytic concepts such as heterogeneity, transitivity, and inconsistency. The proposed framework provides a structured way to formulate and synthesise complex clinical questions in contemporary evidence synthesis.

[28-P6-13] Cross-design evidence synthesis in individual-participant data meta-analysis of treatment effects: a scoping review of methods and recommendations for reporting and conduct

Presenter: Jifan Wang

Authors: Jifan Wang [1]; Demy Idema [1]; Karel Moons [1]; Joost van Rosmalen [1]; Valentijn de Jong [1]

Affiliations: [1] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht

Introduction

Evidence from randomized controlled trials (RCTs) and non-randomized studies (NRSs) is increasingly combined in meta-analyses, leading to cross-design evidence synthesis. Advanced statistical methods may be needed to mitigate biases introduced by cross-design synthesis, especially when using individual participant data. We aimed to review the occurrence and rationale of cross-design synthesis, and the analytical choices in individual participant data meta-analyses (IPDMAs) in clinical research.

Methods

We systematically searched PubMed for IPDMAs that obtained or aimed to include both randomized and non-randomized data to estimate a treatment effect. Two reviewers screened the articles independently. Discrepancies were resolved by discussion or consultation with a third reviewer. Data extraction was performed by one reviewer and checked by the other reviewer.

Results

Among the 75 included IPDMAs, 40 (53%) studies mentioned the rationale for proposing cross-design search or synthesis. The top reason (21%) was insufficient evidence from RCTs on the treatment effect. Fifty-seven studies (76%) identified or utilized evidence from both RCTs and NRSs. The treatments in RCTs were randomized to the target intervention of the IPDMA in only about half of the studies (49%), and this information was unclear in 21 studies (37%). Most studies used one of four analytical approaches: pooling RCTs and NRSs in the primary analysis (75%), separate analyses by study design only (5%), both pooled and separate analyses (9%), or pooling RCTs only (9%). Half of

the studies (51%) applied random effects to adjust for between-study variation. One IPDMA reported adjusting for study design. Another IPDMA tested the treatment-study-design interaction while the reported analyses were separately conducted by RCTs and NRSs. No study used propensity score methods to account for cross-design synthesis, and no other cross-design synthesis method was reported.

Discussion

The reporting of data obtained and statistical methods for performing cross-design synthesis was generally poor. We recommend assessing whether the randomizations in RCTs align with the interventions studied in the IPDMA and reporting the reassessed study design. When the interventions are not aligned, the data should be treated as non-randomized in the analysis. Systematic reviewers should consider the difference in study designs, use appropriate methods to reduce bias in estimates, and clearly report the statistical methods.

[28-P6-14] Synthesizing randomized and non-randomized evidence in meta-analysis: a scoping review of methodologies and a simulation study of the various approaches

Presenter: Ourania Koutsiouroumpa

Authors: Ourania Koutsiouroumpa [1]; Georgios Seitidis [1, 2]; Stavros Nikolakopoulos [2, 3]; Dimitrios Mavridis [1]

Affiliations: [1] Department of Primary Education, University of Ioannina, Ioannina, Greece, Ioannina; [2] Department of Psychology, University of Ioannina, Ioannina, Greece, Ioannina; [3] Department of Data Science & Biostatistics, Julius Center, University Medical Center Utrecht, Utrecht

Introduction

Meta-analyses of Randomized controlled trials (RCTs) are considered the most reliable source of evidence for estimating treatment effects in meta-analysis (MA). However, RCT evidence may be limited when outcomes are rare, sample sizes are small, or follow-up periods are short, affecting external validity. Non-randomized evidence (NRE), such as cohort studies and registry data, can provide complementary information but is more susceptible to confounding and selection bias. Several statistical methods have been proposed to combine randomized and NRE within MA and network meta-analysis (NMA), most commonly within Bayesian hierarchical frameworks.

Methods

We conducted a methodological scoping review of statistical approaches for combining randomized and NRE, searching PubMed, Scopus and Embase. The identified methods were illustrated using a rare-event dataset on methotrexate exposure and melanoma risk. In addition, a simulation study was conducted to compare ten integration strategies under scenarios involving observational bias, population differences, dominance of observational evidence and increased heterogeneity under rare-events settings. Performance metrics included bias, coverage, power, RMSE and credible interval width. We developed R code to facilitate the implementation of the different approaches.

Results

Identified methods were classified in the following categories: naïve pooling, design-adjusted synthesis, prior-based borrowing approaches (e.g., meta-analytic predictive (MAP) priors and robust MAP priors) and hierarchical cross-design models. In the empirical example, treatment effect estimates were broadly consistent across models, while interval width and heterogeneity varied. Preliminary simulations (R=100) showed differences in performance across scenarios. Design-adjusted approaches demonstrated lower bias. Analyses based only on randomized evidence showed greater variability and wider intervals, particularly under vague prior assumptions. Prior-based methods (MAP and robust MAP) produced wider intervals, improving coverage but reducing power. Coverage was close to 95% in effect scenarios (OR≠1) but decreased in null scenarios (OR=1) for models sensitive to biased observational evidence, indicating higher false positive rates.

Discussion

These preliminary results showed that methods that explicitly account for differences between study designs provide more robust inference. Ongoing work with increased simulation repetitions (e.g., 1000) and scenarios (e.g., varying baseline probability) will assess the stability of these results. Future work includes the development of an R package for implementing the proposed models.

[28-P6-15] Bayesian Methods in Registered Clinical Trials: A Systematic Review of Studies on ClinicalTrials.gov through 2024

Presenter: Christina Geyer

Authors: Christina Geyer [1]; Giles Partington [2]

Affiliations: [1] Phastar, Regensburg; [2] Phastar, London

Introduction

Bayesian methods are increasingly being incorporated into clinical trial designs to improve flexibility, efficiency, and interpretability. Earlier reviews of published studies (Lee & Chu, 2012) illustrated how these approaches were applied across selected examples. Building on that foundation, we conducted the first registry-based review to describe how Bayesian methods are currently represented in interventional drug trials, using ClinicalTrials.gov as the primary data source.

Methods

All interventional trials registered on ClinicalTrials.gov through December 2024 were screened for mention of a Bayesian component, using search terms spanning explicit Bayesian terminology and method-specific acronyms (BOIN, BLRM, CRM, MAP, MCMC). After removing non-drug and duplicate records, eligible trials were categorized by therapeutic area, trial size, trial phase, design type, and the role of the Bayesian methods within the analysis framework.

Results

From over 82,000 registered trials screened, 422 interventional drug trials with a Bayesian element were identified (approximately 0.5%). The review shows a steady increase in adoption over time, steepest post-2012. Bayesian methods served as the primary analytic approach in 86% of identified trials, with early-phase oncology dose-finding studies remaining the most frequent application.

Discussion

This work provides the first quantitative snapshot of Bayesian clinical trial activity derived directly from registry data. The steady increase we document across therapeutic areas, phases, and design types reflects genuine broadening of Bayesian adoption beyond early-phase oncology. Our findings establish a current baseline for future comparisons and are timely alongside the recently released FDA draft guidance on Bayesian methods for drugs and biologics (January 2026), providing an account of registered trial activity in the period leading up to that guidance.

[28-P6-16] Small-study effects investigated by Bayesian simulations of trial continuations

Presenter: Markus Huber

Introduction

In meta-analyses, small-study effects refer to the situation where trials with small sample size may exhibit different and/or larger treatment effects than trials with larger sizes, with potential implications for the meta-analyses' validity. Over the past decades, a number of qualitative and quantitative methods were proposed to investigate small-study effects. Here, a Bayesian method is proposed to answer the question: "What would be the pooled effect of a meta-analysis if all trials had the trial size of the largest trial?"

Methods

We used binary outcomes of meta-analyses in the Cochrane Database of Systematic Reviews (CDSR) previously compiled and analysed by Lin et al. (DOI: 10.1136/bmjebm-2019-111191). We considered 2'635 systematic reviews with a total of 43'828 meta-analyses. For each meta-analysis, we simulated a hypothetical continuation of each trial up to the sample size of the largest trial as follows: For each trial, a Beta prior (α=1,β=1) was updated according to the trial's observed outcomes in the control and intervention group. The result is a Beta posterior from which hypothetical, newly enrolled patients can be sampled until the target sample size - that is, the size of the largest trial in a particular meta-analysis - is reached. Each trial was hypothetically continued 100 times and 100 pooled effect sizes based on a random-effects model were calculated.

Results

In 60.7% of the meta-analyses, the median P-value based on 100 trial continuations with equal sample size increased compared to the original P-value. In 4.6%, the median P-value was not statistically significant, in contrast the original, statistically significant P-value, which could be interpreted as a sign of small-study effects. Additionally, in 2.4%, the median P-value was statistically significant,

whereas the original P-value was not statistically significant, which could be interpreted as a sign of underpowered studies.

Discussion

A Bayesian framework provides an intuitive way to study the impact of differing sample sizes of trials on the magnitude and statistical significance of the pooled effect size. Currently at a proof-of-concept stage, further investigations need to assess the method's robustness properties, in particular with respect to the number of required trial continuations.

[28-P6-17] When Does Meta-Regression Work? A Simulation Study on Aggregation Bias in Continuous Covariates

Presenter: Mikolt Bakony

Authors: Mikolt Bakony [1]

Affiliations: [1] Centre for Translational Medicine, Semmelweis University, Budapest, Budapest

Introduction

Meta-regression is commonly used to investigate effect modification using study-level summaries of continuous covariates. However, its validity depends on how covariates vary within and between studies, and may be compromised by Aggregation bias. In applied research, different types of covariates exhibit distinct variability patterns: design-related variables such as treatment dose or pre-specified follow-up time often show limited within-study variation, whereas patient characteristics such as age or body mass index (BMI) typically vary substantially within studies. The implications of these differences for meta-regression are not well characterized.

Methods

We conducted a simulation study comparing aggregate meta-regression with mixed-effects models fitted to individual patient data (IPD). We generated datasets with 10 studies and 100 patients per study under varying ratios of within- and between-study variability for a continuous covariate. Scenarios reflected realistic use cases, including low within-study variability (dose, follow-up time) and high within-study variability (age, BMI). Outcomes were simulated from hierarchical models allowing distinct within- and between-study covariate effects. Covariate effects were estimated using linear mixed models on IPD and random-effects meta-regression based on study-level estimates. Performance was assessed in terms of bias, mean squared error, and coverage across 1000 simulations.

Results

Meta-regression performed well when covariates varied primarily between studies, yielding estimates comparable to IPD analyses. In contrast, for covariates with substantial within-study variability, meta-regression produced biased estimates with poor coverage. When within- and between-study effects differed, aggregate analyses occasionally reversed the direction of the association, consistent with Simpson's paradox.

Discussion

These findings illustrate how the performance of meta-regression depends on the underlying data-generating structure, particularly the relative magnitude of within- and between-study variability in covariates. While aggregate analyses appeared adequate for variables with primarily between-study variation (e.g., treatment dose or follow-up time), they were more prone to bias for covariates with substantial within-study variability, such as age or body mass index. This has practical implications, as such patient-level characteristics are frequently used in applied meta-regression. These observations support a more cautious interpretation of meta-regression results and suggest that, where feasible, individual patient data approaches may provide more reliable estimates of covariate effects.

P7 - AI and Biomedical Data

Poster area: P7

[28-P7-01] How Many Samples Are Enough for Reliable Machine Learning? A Learning Curve-Based Study from SNP Simulations to Alzheimer's Proteomics

Presenter: Nurgül Bulut

Authors: Nurgül Bulut [1]; Pınar Gunel [2]; Zeliha Aydin Kasap [3]; Stéphanie Boutillier [4, 5]; Huseyin Firat [5, 4]; Seval Kul [6, 7]

Affiliations: [1] Istanbul Medeniyet University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, İstanbul; [2] Alanya Alaaddin Keykubat University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Antalya; [3] Karadeniz Technical University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Trabzon; [4] Amoneta Diagnostics SAS, Huningue; [5] Firalis SA, Huningue; [6] Gaziantep University, Faculty of Medicine, Department of Biostatistics, Gaziantep; [7] Cardiovascular Research Institute Basel (CRIB), Basel

Introduction

Determining how many samples are needed to build reliable machine learning (ML) models remains a fundamental question in high-dimensional biomarker research. In omics studies, where predictors far exceed samples, insufficient sample sizes can lead to unstable models and overly optimistic performance estimates. Despite the widespread use of LASSO and XGBoost, practical guidance on sample size requirements remains limited. This study systematically evaluates the impact of sample size, dimensionality, signal strength, and class imbalance on model performance.

Methods

We developed a simulation framework using learning-curve analysis on a virtual population (N=30,000) with correlated SNP-like predictors (ρ=0.30). Dimensionality ranged from p=300 to 10,000 with 1% and 5% defined as true signals calibrated for target AUCs (0.70-0.90). LASSO and XGBoost models were trained on varying bootstrap-sampled sizes under balanced and unbalanced scenarios, with

performance evaluated across 100 replicates using a fixed independent test set (n=1,000). Learning curves were used to assess performance trends and variability as sample size increased.

Finally, the framework was validated using ADDIA (Horizon 2020) project proteomics data (~5,400 proteins) comparing Alzheimer's patients and healthy controls.

Results

Model performance increased with sample size, but convergence strongly depended on dimensionality. For moderate dimensionality (p=300), AUC rapidly approached ~0.70, reaching ~0.68-0.70 with approximately 1,000-1,500 samples and stabilizing thereafter. In contrast, for higher dimensional settings (p=1,000), required ~2,000-3,000 samples for similar levels (AUC~0.65).

In very high-dimensional scenarios (p=5,000-10,000), performance remained substantially lower (AUC~0.52-0.58) with high variability, even at the largest sample sizes, indicating limited signal recovery. LASSO was more stable than XGBoost in low-sample regimes.

Importantly, these patterns were confirmed by ADDIA proteomics data (p≈5,400), showing delayed stabilization and persistent variability, consistent with simulation results.

Discussion

This study provides a practical answer to the question of how many samples are needed in high-dimensional ML-based biomarker studies. Results show that required sample size scale sharply with dimensionality and lower signal strength, highlighting early performance estimates can be misleading. Robustness was confirmed by the consistency between SNP-based simulations and real-world Alzheimer's proteomics data. Practically, we recommend incorporating learning-curve analyses during study design and avoid drawing conclusions from small-sample ML models, in high-dimensional omics settings.

[28-P7-02] Statistical models for surgical learning curves

Presenter: Marianne Huebner

Authors: Marianne Huebner [1]; Rayleigh Lei [1]; Wenjuan Ma [1]

Affiliations: [1] Michigan State University, East Lansing

Introduction

Learning curves graphically represent a physician's skill over time, helping identify the point at which proficiency is achieved. Outcomes may be continuous, such as operative time, or discrete, such as the occurrence of major adverse cardiovascular events (MACE). Several statistical approaches have been used to develop surgical learning curves, including cumulative summation, regression models, discrete interval approaches (e.g., every 10 cases), and, more recently, machine learning models. However, these approaches typically do not account for patient characteristics, may disregard the correlation structure inherent to the longitudinal nature of physician-performed procedures, and often fail to examine underlying data properties, potentially leading to inappropriate exclusion criteria for learning curve analysis.

Methods

Using the example of a minimally invasive procedure for atherosclerotic carotid artery disease we discuss two approaches for analyzing learning curves: Bayesian regression and neural networks. For the neural networks method, we systematically selected from several distinct architectures, including Transformer, Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and hybrid CNN-GRU models. After identifying the best-performing architecture, we conducted hyperparameter tuning to optimize model performance. We then performed an ablation study to evaluate variable importance and identify the key features influencing outcomes.

Results

Initial data analyses led to further exclusion criteria. As the number of procedures increased, surgeon performance improved, demonstrating a clear learning effect until approximately 150-200 procedures. Incorporating patient variables into the models revealed that the observed decrease in operative time and adverse events may instead be driven by other factors, such as case urgency. The ablation study identified the most important features for predicting outcomes including insurance type and emergency status.

Discussion

These findings underscore that approaches to estimate surgical learning curves can lead to significant over- or underestimation when patient characteristics are not considered. The application of neural network architectures, coupled with rigorous feature importance analysis via ablation studies, provides a more nuanced understanding of the factors driving surgical performance and highlights the necessity of accounting for clinical context in learning curve assessments.

[28-P7-03] Comparison of Statistical and Neural Sequential Approaches for Early Disease Detection in Claims-Based Healthcare Event Sequences: Assessing the Role of Event Order and Timing

Presenter: Corentin Faujour

Authors: Corentin Faujour [1, 2]; Stéphane Bouee [1]; Corinne Emery [1]; Anne-Sophie Jannot [2]

Affiliations: [1] CEMKA, Bourg-La-Reine; [2] HeKa Université Paris Cité, INRIA, INSERM, F-75015, Paris

Introduction

Claims databases can provide longitudinal representations of healthcare trajectories that can be used for early disease detection, particularly in the context of diagnostic delay. While standard statistical approaches rely on aggregated features that ignore event order and timing, neural sequential models (LSTM, Transformer) account for event order, yet the respective contributions of event order and temporal information remain unclear. We evaluated their added value for early classification of patient trajectories in terms of predictive performance and interpretability.

Methods

Healthcare event sequences were simulated for 50,000 cases and 50,000 controls using parameters estimated from 22,000 amyotrophic lateral sclerosis cases and 22,000 birth year and sex-matched controls from the French National Health Data System (SNDS).

Events were generated using a time-varying process distinguishing background events (non-specific healthcare activity) and condition-related events. In cases, the frequency of condition-related events increased as diagnosis approached.

Three model categories were compared: (i) statistical models based on aggregated event frequencies (logistic regression, SVM), (ii) neural sequential models accounting for event order, and (iii) neural sequential models incorporating temporal information (absolute time or inter-event delays).

Performance was evaluated using the area under the ROC curve (AUC) at multiple time points before diagnosis, with confidence intervals estimated using the DeLong method.

Results

Frequency-based models achieved an average AUC of 0.70 (95%CI: 0.68-0.72). Neural sequential models showed improved predictive performance, with a consistent gain of +0.06 AUC across all time horizons, reflecting the contribution of event order. In contrast, incorporating inter-event delays did not provide consistent improvement.

Regarding interpretability, both neural sequential models and logistic regression correctly identified condition-related events defined in the simulation as the main contributors to prediction. Neural sequential models offer the additional advantage of providing a dynamic representation of event contributions along the trajectory, highlighting time periods where predictive signal is strongest.

Discussion

Modelling event order improves classification of healthcare trajectories, while time information provides limited additional value. Neural sequential models showed consistency with statistical approaches in terms of identified signals, supporting their use as an interpretable extension in this setting.

[28-P7-04] Using Survival Neural Networks on Synthetically Created Data with known Ground Truth.

Presenter: Benjamin Keane

Authors: Benjamin Keane [1]; Jérémie Nsengimana [2]; Jingky Lozano-Kuehne [3]

Affiliations: [1] Newcastle University, Newcastle upon Tyne; [2] Population Health Sciences Institute, Newcastle University, Newcastle upon Tyne; [3] Population Health Sciences Institute, Newcastle University, UK, Newcastle upon Tyne

Introduction

Developing prognostic and predictive biomarkers is a key step in precision medicine. We are conducting a simulation study of large scale -omics in cohorts of cancer patients to assess the strengths and limitations of deep learning-based methods in predicting patient outcomes with high-dimensional data. Biology is complex and difficult to simulate, very few simulators of -omics data exist that preserve complex networks of correlations and interactions at genome-wide level. Umpire 2.0 package in R Bioconductor is one of these tools, but it has received, to our knowledge, limited independent evaluation. Our study has two aims: a) Assessing the capabilities of this algorithm in producing realistic transcriptomic profiles linked to metadata, including survival b) using these data with known ground truth to test state-of-the-art Artificial Neural Networks (ANN) and Random Survival Forest (RSF) methods to predict patient outcomes. Evaluating performances of these methods will allow identifying areas of possible improvements.

Methods

We used the Umpire 2.0 package to simulate transcriptomes of 704 patients, grouped in 220 biochemical pathways (average 100 genes per pathway), including 20 impacting survival with a median beta coefficient 0.9 (range 0.06-2.7) in Cox proportional hazards models. ANN analysis was conducted within Pycox package1 and RSF within scikit-survival package2, both in python. Models were compared by C-index and Brier scores.

Results

Preliminary results indicate that umpire 2.0 simulated good quality transcriptomes with simple disease models (no interactions) and realistic metadata. Patient survival was predicted with a C-index up to 0.66 [95% CI 0.61-0.71] by ANN and 0.61 [0.49-0.73] by RSF, Brier score 0.14 [0.12, 0.16] (ANN) and 0.12 [0.09-0.14] (RSF).

Discussion

In these limited analyses ANN achieves a better C-index with a comparable brier score to RSF. This is a first run of two prediction tools with default parameters. We are refining the simulation process and analysis models and will present in the conference more extensive comparisons.

1. Kvamme, H., Borgan, O. & Scheel, I. Time-to-event prediction with neural networks and cox regression. Journal of Machine Learning Research 20, (2019).

2. Pölsterl, S. Scikit-survival: A library for time-to-event analysis built on top of scikit-learn. Journal of Machine Learning Research 21, (2020).

[28-P7-05] Age Estimation using Multiple Instance Learning based on Ultrasound Images of the Knee

Presenter: Marco Pawłowski

Authors: Marco Pawłowski [1]; Eilin Jopp-van Well [2]; Jochen Herrmann [2]; Heinz Handels [3]; Dennis Säring [1]

Affiliations: [1] Fachhochschule Wedel gGmbH, Wedel; [2] Universitätsklinikum Hamburg-Eppendorf, Hamburg; [3] Institute of Medical Informatics, University of Lübeck, Lübeck

Introduction

Accurate age determination is crucial in forensic medicine, particularly for individuals without valid identification. Epiphyseal ossification of the knee joint, strongly correlates with chronological age and serves as a relevant anatomical marker. While MRI enables reliable non-invasive estimation, its cost and limited accessibility restrict use. Ultrasound provides a cost-effective and portable alternative, with prior studies showing comparable manual results to MRI [1]. CNN-based analysis of MRI slices has achieved high performance (accuracy 0.924, F1 score 0.941), highlighting the potential of automation [2]. We therefore explore automated age estimation from ultrasound using a multiple instance learning (MIL) framework.

Methods

A dataset of 20 male subjects aged 14-20 years was analyzed, comprising paired ultrasound and MRI data acquired in a Deutsche Forschungsgemeinschaft (DFG) funded clinical trial. Ultrasound data were recorded as time series with labels available at the sequence level, while not all frames depicted relevant anatomical structures. To address this, we employ a MIL framework, modeling each ultrasound sequence as a bag of instances under weak supervision. The approach accommodates variable-length sequences and produces a binary age classification (=18 years old). A neural network-based, permutation-invariant aggregation operator, implemented via an attention mechanism, assigns weights to frames [3].

Results

The proposed method attains an accuracy of 0.857 and an F1 score of 0.667, while providing interpretability via the MIL framework, with absolute differences of -0.067 accuracy and -0.274 F1 score relative to MRI-based CNN approaches.

Discussion

Performance differences relative to MRI-based CNN approaches likely arise from ultrasound-specific challenges, including variability in sequence length, weak supervision from sequence-level labels, and data sparsity, as each subject contributes a single sequence versus multiple annotated MRI slices. Nevertheless, results are promising and require further evaluation.

- [1] Hermann et al., Forensic age assessment of the knee: proposal of a new classification system using two-dimensional ultrasound volumes and comparison to MRI

- [2] Pawłowski et al., Comparison of Modern Transformer Architectures and CNN-Based Models for MRI-Based Age Estimation of the Knee

- [3] Isle et al., Attention-based Deep Multiple Instance Learning

[28-P7-06] Explainable Graph Neural Networks for detecting myocardial infarction from 12-lead ECGs

Presenter: Miriam Cindy Maurer

Authors: Miriam Cindy Maurer [1, 2]; Philip Hempel [1]; Kristin Steinhaus [3]; Hryhorii Chereda [4]; Marcus Vollmer [5, 6]; Dagmar Krefting [1, 2]; Nicolai Spicher [7]; Anne-Christin Hauschild [8]

Affiliations: [1] University Medical Center Göttingen, Department of Medical Informatics, Göttingen; [2] Campus Institute Data Science (CIDAS), Universität Göttingen, Göttingen; [3] Clinic of Cardiology and Pneumology, Göttingen, Göttingen; [4] Biomedical Network Science Lab, Department of Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander University Erlangen-Nürnberg, Erlangen; [5] Universitätsmedizin Greifswald, Greifswald; [6] Deutsches Zentrum für Herz-Kreislauf-Forschung e.V. (DZHK), Greifswald; [7] Technical University of Denmark, Kongens Lygnby; [8] Institute for Predictive Deep Learning in Medicine and Healthcare, Gießen

Introduction

Artificial intelligence (AI) is becoming more frequently used in electrocardiography (ECG) for the detection of cardiovascular diseases. Despite the high predictive performance, the lack of interpretability and transparency limits translation towards clinical application. Novel technologies, such as Graph Neural Networks (GNNs) offer a natural way to model clinically relevant spatial and temporal relationships between ECG leads. To address this, integrating explainable AI (XAI) is crucial to ensure transparency, trustworthiness, and compliance with emerging regulations such as the EU AI Act. Therefore, this work introduces an explainable GNN framework for clinically meaningful ECG analysis with a focus on myocardial infarction (MI).

Methods

We developed xGNN4MI [1], an open-source pipeline for 12-lead ECG classification based on graph representations. ECG signals are transformed into graphs that capture inter-lead dependencies and temporal structures. A Graph Convolutional Network is trained on the PTB-XL dataset for diagnostic superclass classification and MI localization. The explainability of the model is achieved by integrating GNNExplainer, which identifies the relevant leads and connections that contribute to the model's decisions. The generalizability of the results is assessed using the external SHIP cohort.

Results

The model achieved a good performance for superclass classification (AUC = 0.86, F1 = 0.68) and improved results for MI localization (AUC = 0.92, F1 = 0.78). External validation demonstrated good generalizability (AUC = 0.86, F1 = 0.68). Explainability analyses revealed clinically consistent patterns: anterior leads (V1-V3) were most relevant for anteroseptal MI, while inferior leads (II, III, aVF) dominated for inferior MI. In contrast, healthy ECGs showed diffuse relevance patterns across leads.

Discussion

xGNN4MI provides an open-source, reproducible, and explainable framework for ECG-based disease classification. The alignment of model explanations with established cardiological knowledge supports its clinical plausibility and trustworthiness. By combining GNNs with systematic XAI evaluation and full code availability, this work contributes to transparent and reproducible AI in medicine and supports future research and clinical translation.

[1] Maurer, M.C. et al. xGNN4MI: explainability of graph neural networks in 12-lead electrocardiography for cardiovascular disease classification. npj Digit. Med. (2026). https://doi.org/10.1038/s41746-026-02367-1

[28-P7-07] Multilayer Modeling of Brain Hierarchy for Interpretable Disease-Association Analysis Using Neuroimaging Data

Presenter: Suzuna Nishi

Authors: Suzuna Nishi [1]; Ryo Emoto [2]; Kota Matsui [2]; Shigeyuki Matsui [2]

Affiliations: [1] Department of Clinical Biostatistics, Kyoto University School of Public Health, Kyoto; [2] Department of Biostatistics and Data Science, Kyoto University School of Public Health, Kyoto

Introduction

The analysis of neuroimaging data to identify brain regions associated with disease is essential for understanding disease pathology and generating biologically meaningful interpretations. However, there is a gap between many disease-association analysis and their scientific interpretations: while statistical analyses are often performed at the voxel level, scientific interpretations are made at the level of functional regions or large-scale networks. As a technical issue, since conventional spatial models mainly account for local continuity, the common practice of subjective post-hoc aggregation of significant voxels may prevent a formal Type I error control for the identified functional regions. To bridge the aforementioned gap, this study proposes a multilayer stochastic model that incorporates the structural hierarchy of layers of voxels, functional regions, and large-scale networks.

Methods

We formulate a hierarchical probabilistic model for the observed effect sizes and introduce latent variables that represent disease associations at the voxel, region, and network layer levels. By incorporating this hierarchical structure, the proposed model captures the dependency induced by the anatomical and functional hierarchy within a unified inference framework. Based on posterior quantities derived from this model, we construct a within-layer multiple testing procedure to control the false discovery rate (FDR) at both the voxel and functional-region layers.

Results

In simulations, the proposed method maintained the voxel-level FDR below the target level and achieved power close to that of an oracle procedure assuming known model parameters. At the

region level, the method showed FDR inflation under weak signals in some settings, but this behavior diminished as signal strength increased, with region-level performance approaching that of the oracle method. We further applied the method to neuroimaging data from the ABIDE study in Autism Spectrum Disorder.

Discussion

Our approach provides a principled alternative to subjective post-hoc aggregation for interpreting neuroimaging associations beyond isolated voxels. In particular, it enables formal FDR-controlled inference at the voxel level and under sufficient signal strength, at the region level as well. The region-level inflation observed under weak signals suggests a phase-transition-like regime in multilayer inference, likely reflecting propagation of voxel-level noise to upper layers. Improving robustness in this setting remains an important direction for future work.

[28-P7-08] Simultaneous detection of metastases and classification of underlying primary tumors in cranial MRI scans using Deep Learning

Presenter: Christian Schorr

Authors: Christian Schorr [1]; Wolfgang Reith [2]; Malvina Garner [2]; Sebastian Hocke [1]; Aseer Ansari [1]; Umut Yilmaz [2]

Affiliations: [1] German Research Center for Artificial Intelligence (DFKI), Saarbrücken; [2] Department of Neuroradiology, Saarland University, Homburg (Saar)

Introduction

Rapid detection of brain metastases and early identification of the most likely primary tumor are clinically important, as both directly influence further diagnostic workup and treatment planning. An AI-based copilot that detects metastases and simultaneously suggests the underlying primary malignancy on routine contrast-enhanced 3D T1-weighted MRI could therefore support radiologists in daily workflow.

Methods

Our data set contained 157 patients with one or more brain metastases. Of those 36 were caused by melanoma, 21 by breast cancer and 100 by lung cancer. MRI data consisted of contrast-enhanced 3D T1-weighted sequences (T1-MPRAGE or T1-SPACE). Across these examinations, 4.397 slices contained metastases, including 682 from breast cancer, 1.258 from melanoma, and 2.457 from lung cancer. We also added 4.395 slices without metastases to the data set.

We trained the object detection algorithm YOLO 26n for 100 epochs with light image augmentation to simultaneously detect metastases and classify their underlying primary tumour. The data was split into 80% training, 10% validation and 10% testing data while keeping the class proportions constant.

To improve performance we also implemented a multi-orientation approach, calculating coronal and axial slices from the 3D MRI scan and training separate YOLO models for these orientations. The predictions of all three models were then combined through a voting strategy to reduce false negative and false positive predictions.

Results

The sagittal model was able to recall 90.9% of all metastases with a precision of 89.6%. The metrics MAP50 = 0.909 and MAP50-95 = 0.746 demonstrate a good robustness. An analysis of the confusion matrix shows that metastases caused by lung and breasts cancers are deteced very reliably while melanomas are sometimes missed. Coronal and axial model metrics are slighty lower than for the sagittal model. The combination step of the proposed multi-orientation approach is still being implemented.

Discussion

Deep learning applied to routine contrast-enhanced 3D T1-weighted MRI enables reliable detection of brain metastases and shows promising performance for simultaneous classification of the underlying primary tumor. The planned integration of sagittal, axial, and coronal models may further improve robustness by combining complementary detection and classification of the same lesion.

[28-P7-09] When More Data Fails: A Study of Domain Shift in DL-Based Prostate Cancer Segmentation

Presenter: Jakub Mikolajczak

Authors: Jakub Mikolajczak [1, 2]; Alina Szabo [2]; Daniel Schmidt [2]; Dirk Hellwig [2]

Affiliations: [1] Medizinisches Datenintegrationszentrum am Universitätsklinikum Regensburg, Regensburg; [2] Universitätsklinikum Regensburg, Regensburg

Introduction

Prostate cancer is one of the most frequently diagnosed malignancies among men worldwide, with approximately 1.5 million new cases annually. Advanced and metastatic prostate cancer remains associated with poor prognosis [1].

PSMA PET/CT plays a central role in diagnosis and treatment planning. In this context, image-based biomarkers (IBBs) are increasingly used for quantitative disease assessment. However, their computation requires manual annotation, which is time-consuming and prone to inter-observer variability.

Automated deep learning-based segmentation offers a promising solution to enable efficient and reproducible extraction of IBBs.

Methods

We trained three convolutional neural networks (CNNs) models using nn-UNet framework [2] on three different datasets:

1. 68Ga PSMA subset from AutoPET dataset [3]

2. whole PSMA subset from AutoPET dataset

3. PSMA subset from AutoPET + DEEP-PSMA dataset [4]

Results

For the three different models we obtained following Dice Scores: 0.66, 068, 0.54.

Discussion

Unexpectedly, adding the DEEP-PSMA dataset led to a substantial drop in performance, despite increasing the amount of training data. This degradation is mainly driven by a strong increase in false positive predictions, indicating reduced specificity while sensitivity remains similar.

A likely explanation is the difference in intensity distributions between datasets. DEEP-PSMA contains a higher proportion of tumor voxels with lower SUV values, which may cause the model to classify lower-intensity regions as malignant more frequently, leading to over-segmentation.

These results suggest that simply increasing dataset size does not guarantee improved performance. Instead, domain shifts between datasets must be carefully addressed, for example through domain adaptation or intensity normalization.

References:

1. Raychaudhuri R, Lin DW, Montgomery RB. Prostate Cancer: A Review. JAMA. 2025333(16):1433-1446. doi:10.1001/jama.2025.0228

2. Isensee, F., Jaeger, P.F., Kohl, S.A.A. et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 18, 203-211 (2021). https://doi.org/10.1038/s41592-020-01008-z

3. Gatidis S, Früh M, Fabritius MP, et al. Results from the autoPET challenge on fully automated lesion segmentation in oncologic PET/CT imaging. Nat Mach Intell. 20246:1396-1405. doi:10.1038/s42256-024-00912-9.

4. Price J, Hofman M, Buteau JP, McIntosh L, Sun Y. A repository of annotated PSMA and FDG PET/CT images for algorithm development in staging of mCRPC for treatment with 177Lu-PSMA therapy [dataset]. Zenodo 2025. doi:10.5281/zenodo.15281784

Tue 29.09.2026

P1 - Digital Health and Data Systems

Poster area: P1

[29-P1-01] Visualizing Data Quality Results - Status Quo and New Solutions

Presenter: Stephan Struckmann

Authors: Stephan Struckmann [1]; Dennis Hübner [2]; Carsten Oliver Schmidt [3]

Affiliations: [1] Universitätsmedizin Greifswald, Greifswald; [2] Thiem Research GmbH, Cottbus; [3] Universität Greifswald, Greifswald

Introduction

Data quality assessments generate many findings across aspects such as data completeness, consistency, outliers, variable distributions and associations [1,2]. In large studies, this makes it difficult to obtain an overview without losing detail. The challenge is visual and methodological, because summarisation requires an integration algorithm that combines findings across indicators, variables, and examinations without obscuring critical problems. We therefore reviewed current approaches to displaying integrated data quality results and developed an interactive hierarchical visualisation for epidemiological data quality reports.

Methods

The work had two components. First, we conducted a narrative review of papers and software identified via PubMed and Google Scholar, focusing on visualisation and aggregation of data quality findings in software tools, dashboards, and recent EHDS-related work [2]. Second, we implemented a hierarchical sunburst visualisation in the R package dataquieR and applied it to data from the population-based Study of Health in Pomerania (SHIP) [3]. The implementation maps findings to a hierarchical framework [1] and supports two integration modes: a conservative non-additive approach, in which single critical findings dominate parent segments, and a compensatory additive approach based on weighted aggregation. Users can adjust between these modes.

Results

The review identified more than 15 heterogeneous concepts and tools. Integrated cross-dimensional summarisation remains uncommon. Most tools either display many checks side by side or summarise only within dimensions, such as completeness or plausibility. In contrast, additive approaches such as the EHDS/QUANTUM label derive an overall score, whereas conservative approaches allow critical findings to dominate. To reconcile both, sunburst charts were integrated into dataquieR. In the SHIP

example, this enabled compact display of multi-level results and made the consequences of different integration rules transparent.

Discussion

The appropriate choice of an integration approach depends on the analytical objective, such as obtaining an overall assessment of a dataset versus identifying specific patterns of data quality problems. Making aggregation logic explicit and adjustable improves transparency and supports context-specific interpretation. Hierarchical sunburst charts offer an effective way to combine overview and in depth-insights in complex data quality assessments.

References

[1] https://doi.org/10.1186/s12874-021-01252-7

[2] https://tehdas.eu/app/uploads/2022/05/tehdas-european-health-data-space-data-quality-framework-2022-05-18.pdf

[3] https://doi.org/10.1093/jamia/ocab132

[29-P1-02] An Iterative Data Quality Framework for an International Clinician-Reported Rare Disease Registry: Experience from the ICI-Myocarditis Registry

Presenter: Shuang Xia

Authors: Shuang Xia [1]; Adrien Procureur [1]; Stephane Ederhy [1]; Joe-Elie Salem [1]

Affiliations: [1] Sorbonne University, Assistance Publique-Hôpitaux de Paris, INSERM, CIC-2503, Department of Pharmacology, Groupe Hospitalier Pitié Salpêtrière, Paris, France., Paris

Introduction

International multi-center registries are essential for studying rare adverse events such as ICI-related myocarditis. Clinician-reported survey data present unique data quality (DQ) challenges not encountered in electronic health records. Existing frameworks - including the STRATOS IDA Initiative and IEEE Standard 2801-2022 - provide conceptual guidance but lack implementations for voluntary REDCap-based registries. We describe a systematic DQ framework for the ICI-Myocarditis Registry and report on its impact.

Methods

The registry includes 1,255 cases from 168 institutions across 18 countries. Our framework comprises four layers: (1) Metadata-driven data management: automated REDCap data dictionary parsing for variable structure handling, unit standardisation, and missingness profiling. (2) Automated quality checks: 48 programmatic checks in R across four IEEE 2801-2022-adapted dimensions - Completeness (per-section missingness), Accuracy (clinically-informed bounds for ECG, echocardiographic, and laboratory values), Consistency (cross-variable logic, e.g., biomarker initial > peak), and Timeliness (temporal ordering of ICI treatment and follow-up sequences). Each check is tagged to its REDCap section for dimension-stratified reporting. (3) Collaborator feedback loop: issues compiled

into per-collaborator checklists addressing diagnostic eligibility and section-level completeness. (4) Survey instrument evolution: recurring free-text patterns converted into structured response options.

Results

At baseline (September 2025), the pipeline identified 22,018 DQ issues (mean 17.54/case 100% of cases affected). By March 2026 - following one feedback cycle - issues declined to 18,966 (-13.9%), with mean issues/case falling to 15.11 0.4% of cases reached full resolution. Completeness dominated (17,550 issues 79.7%), followed by Accuracy (5.3%), Consistency (1.7%), and Timeliness (0.4%). All dimensions improved: Timeliness (-53.1%), Consistency (-37.6%), Accuracy (-36.8%), and Completeness (-10.3%). Largest section-level reductions: Labs Obtained (-617), Cross-Section Accuracy (-420), Medical Treatments (-372), and Anticancer Drug Coding (-319).

Discussion

A structured, IEEE 2801-2022-grounded DQ framework with automated checks, iterative collaborator engagement, and survey instrument evolution can substantially improve data quality and analytic readiness in large international rare disease registries. This approach is transferable to other REDCap-based multi-center studies and provides a replicable model for pre-analytic data governance.

[29-P1-03] Bringing Transparency to Study Progress: A Distributed Reporting Approach in INTERPOLAR

Presenter: Tatjana Beppler

Authors: Tatjana Beppler [1]; Alexander Strübing [2]; Florian Schmidt [2]; Daniel Neumann [2]; Helene Köster [3]; Sebastian Stäubert [2]; Miriam Kesselmeier [1]; Markus Löffler [2]; André Scherag [1]

Affiliations: [1] Institute of Medical Statistics, Computer and Data Sciences (IMSID), Jena University Hospital, Jena; [2] Institute for Medical Informatics, Statistics and Epidemiology (IMISE), Leipzig University, Leipzig; [3] Medical Centre for Information and Communication Technology, Erlangen University Hospital, Erlangen

Introduction

The INTERPOLAR study aims to implement and evaluate IT support for ward pharmacists during the documentation and identification of medication-related problems (MRPs) in routine clinical care. Using a stepped-wedge cluster randomised design, the study is conducted across 14 German university hospitals as a use case of the MII. It requires an interoperable data integration pipeline combining routine and study data for research purposes.

To support implementation and rollout, a study progress monitoring system is essential, capable of retrieving site-specific insights and creating reports for central discussion and workflow optimisation.

Methods

The INTERPOLAR reporting tool uses distributed analysis to enable a multi-site rollout while keeping data local. Individual case data are aggregated into counts, thereby ensuring compliance with privacy requirements.

Flattened data tables from routine FHIR data are merged with pharmacist documentation and processed in R to calculate reporting metrics. Included as a submodule within the CDS tool chain, scripts can be executed via a single Docker command.

An HTML output ensures readability and standardized export for cross-site aggregation. It enables on-site report visualisation and review by pharmacists and data integration centres (DIC), ensuring data correctness and privacy compliance before transfer to a central, secure cloud.

Results

The distributed reporting approach demonstrates a functioning data pipeline implemented across study sites. Metrics include measures for monitoring study progress and quality, such as:

identification and counting of the study cohort usual care documentation, i.e. quantities of medication analyses and MRPs counts of algorithm-detected MRPs to assess IT-supported workflows and availability of MII broad consent. Local data tables support detection of unexpected data patterns and facilitate debugging. Continuous report development enables early identification of issues during the ongoing study. Between-site and between-ward comparisons provide insights into heterogeneity and challenges in the multi-site analysis pipeline.

Discussion

The approach enables real-time monitoring of recruitment, documentation progress, data quality, and data extraction performance. Importantly, it also enables evaluation of assumptions for statistical analysis and assessment of data availability for the feasibility of secondary analyses. By enhancing transparency and multidisciplinary communication, it builds a framework for improving real-world data quality during the early study stages.

[29-P1-04] Automated generation of clinical trial results based on the CDISC Analysis Results Standard (ARS) using SAS (or R)

Presenter: Wilmar Igl

Authors: Wilmar Igl [1]; Astghik Mikayelyan [2]

Affiliations: [1] VERANEX, Uppsala; [2] VERANEX, Kyiv

Introduction

Data management and statistical analysis consume considerable resources in clinical trials. The high degree of standardization of datasets (ie, SDTM, ADaM) and results (ie, ARS) by the CDISC consortium creates opportunities for automation of these activities. Here, the basic concepts of these

standards and corresponding processes will be explained. Insights from a pilot study for the automated generation of clinical trial results using SAS (or R) will be presented.

Methods

Analysis datasets from clinical studies are highly standardized in the ADaM (Analysis Data Model) model by the CDISC consortium. Recently, also standards for clinical trial analysis results, i.e., CDISC/ARS (Analysis Results Standard), were released (Busa, 2023).

Assuming the combination of ADaM and ARS specifications already determines the statistical analysis required to transform datasets to outputs, the following workflow can be used: 1) Design a Table Template for the expected results, ie, without empirical results. 2.a) Export the Analysis Results Meta-data (ARM) as the abstract description of the table structure and 2.b) export the Analysis Display Meta-Data (ADM) as the concrete description of their graphical layout. 3) Generate the Analysis Programs (ie, SAS/R) based on ADaM and ARM specifications. 4) Generate the Analysis Results Data (ARD), ie, the empirical results, by running the analysis programs. 4) Generate the Final Outputs, eg, PDFs, combining the Analysis Results Data (ARD) and the Analysis Display Meta-Data (AMD), thereby, showing the results in a human-readable format.

Results

A pilot study will be presented using public-use ADaM datasets and a SAS-based framework (Wachara, 2025) to create an automated workflow for the generation of clinical trial results in a business environment. Alternative approaches, e.g., using an R-based framework (Bosman, 2025) including the R/siera and R/tfrmt packages, will be compared. Practical experiences and recommendations will be shared.

Discussion

Advancements in the standardization of individual-level data and group-level results have resulted in automated workflows for the generation of clinical results, which promise efficacy gains for data management and statistical analysis.

References

Busa, B. et al. (2023). PharmaSUG 2023, Paper MM-327.

https://pharmasug.org/proceedings/2023/MM/PharmaSUG-2023-MM-327.pdf Wachara, J. (2025). Phuse US, 2025,

https://www.lexjansen.com/phuse-us/2025/sm/PAP_SM13.pdf Bosman, M. (2025). Phuse EU, 2025, Paper OS20.

https://phuse.s3.eu-central-1.amazonaws.com/Archive/2025/Connect/US/Orlando/PAP_OS20.pdf

[29-P1-05] From Data to Publication: A Reusable SAS Pipeline for Generating and Validating Figures and Tables

Presenter: Louisa Marie Hogger

Authors: Louisa Marie Hogger [1, 2]; Nele Henrike Thomas [1]; Cornelia Frömke [2]; Udo Bavendiek [3]; Johann Bauersachs [3]; Anika Großhennig [1]

Affiliations: [1] Institute for Biostatistics, Hannover Medical School, Hannover; [2] Hochschule Hannover - University of Applied Sciences and Arts, Hannover; [3] Department of Cardiology and Angiology, Hannover Medical School, Hannover

Introduction

The DIGIT-HF trial, a double-blind, randomized, placebo-controlled study evaluated digitoxin's effect on all-cause mortality and hospitalization in heart failure patients with reduced ejection fraction, utilizing a Cox regression model [1]. To prepare and support the primary publication in the New England Journal of Medicine (NEJM) and the first congress presentation, results needed to be formatted into specific tables and figures. To ensure rapid and reliable generation of these outputs after database closure, validated SAS programs were developed beforehand.

Methods

Programs were developed using SAS 9.4 and SAS Macro Language, guided by NEJM requirements, comparable published articles and feedback from other statisticians. A two-step validation process was established. Primary validation used a comparable trial's publication. To confirm functionality across software updates an automated testing program was created using the SASUnit framework [2]. SASUnit, a free testing environment, requires the programming of multiple test scenarios using pre-programmed macros to verify program integrity.

Results

Three distinct SAS programs comprising four different macros were created. The first program generates endpoint tables containing number of events, event rates, hazard ratios with 95%-CI and p-Values, automatically deleting patients with missing relevant variables to ensure consistency. The second program creates cumulative incidence plots, formatting data automatically and displaying the plot with numbers at risk and an optional zoomed inset. The third program produces forest plots for subgroup analysis, handling graphical elements like column positioning and subcategory indentations automatically. All programs support univariate or multivariate analysis for two-arm trials.

Validation via SASUnit involved programming multiple test scenarios to verify table creation and data correctness for both analysis types. Tests require the SASUnit folder structure and files, with results viewable in an HTML report after execution.

Discussion

These programs facilitate the quick creation of tables and figures for two-arm trials with survival endpoints, like DIGIT-HF. Future improvements could include handling of missing data, adding p-values or non-inferiority boundaries and extending the modular design to alternative models like logistic regression.

References

[1] Bavendiek U et al. New England Journal of Medicine. 2025. Available from: https://doi.org/10.1056/nejmoa2415471

[2] HMS Analytical Software. SASUnit. Available from: https://github.com/HMS-Analytical-Software/SASUnit

[29-P1-06] Towards Transparent and Automated Governance of Biomedical Data Use Agreements

Presenter: Jendrik Richter

Authors: Friederike Seyderhelm [1, 2]; Jendrik Richter [1]; Sabine Hanß [1]; Dagmar Krefting [1, 2]

Affiliations: [1] University Medical Center Göttingen, Department of Medical Informatics, Göttingen; [2] Campus Institute Data Science, Georg-August-University Göttingen, Göttingen

Introduction

Biomedical datasets are often governed by complex Data Use Agreements (DUAs), which are needed for ensuring privacy, ethical research conduct, and legal compliance. However, metadata frequently omits the relevant legal details, which are usually embedded within separate usage and/or legal agreements. Moreover, traditional DUAs are typically issued as free-text documents, making them difficult to interpret programmatically therefore limiting automated enforcement. Building on the structured metadata approach proposed by Bönisch et al. [1], this work aims to improve the machine-readability, usability, and enforceability of DUA representations.

Methods

Building on the structured DUA metadata model proposed, we developed a hierarchical, human-centered workflow. DUA clauses are organized into functional groups and descriptive semantic categories, replacing implicit letter codes with interpretable labels. A YAML-based template serves as the single source of truth and is validated and converted into a structured JSON representation, enabling consistent, machine-readable and human-understandable documentation of dataset usage constraints.

Results

The proposed workflow extends prior structured DUA models by introducing a step towards unified and human-centered representation. It enables consistent organization of heterogeneous constraint types, including binary, multi-valued, temporal, and free-text obligations, within a single coherent schema. This improves interpretability and usability while providing a standardized format for reproducible documentation of complex data governance requirements across biomedical datasets. When distributed with datasets, this representation can enhance researchers understanding of data constraints and support automation of associated legal obligations.

Discussion

While the proposed pipeline effectively captures and enforces explicit, rule-based constraints in DUAs, more complex institutional agreements such as SHIP introduce process-dependent elements (e.g., project approval workflows) that cannot be fully represented as static policy rules. This highlights a limitation of purely rule-based formalization approaches and suggests the need for integrating workflow-aware or state-based governance mechanisms. This workflow enables automated

compliance, ensures reproducibility, and improves usability for researchers, creating a scalable and reliable approach to utilizing DUAs for biomedical datasets.

[1] Bönisch C, Hanß, Sabine, Spicher N, Sax U, Krefting D. Reusing Biomedical Data as Agreed - Towards Structured Metadata for Data Use Agreements. In: German Medical Data Sciences 2023 - Science. Close to People. IOS Press 2023. p. 31-8.

[29-P1-07] Assuring trust within the structures of the German National Cohort (NAKO): Concept and implementation of a trusted third party

Presenter: Robert Wolff

Authors: Robert Wolff [1]; Wolfgang Hoffmann [2]; Gunthard Stübs [2]

Affiliations: [1] Trusted Third Party of the University Medicine Greifswald, Greifswald, Germany, Greifswald; [2] Universitätsmedizin Greifswald KöR, Institut für Community Medicine, Greifswald

Introduction

The German National Cohort (NAKO) is the largest population-based epidemiologic cohort study in Germany. Following German and European data protection regulations, in the NAKO a trusted third party (TTP) manages personally identifiable data (IDAT) separately from medical data (MDAT). This work examines the TTP's integration in the NAKO data infrastructure to ensure participant consent, data security, and the right to withdraw - essential for maintaining research integrity and participant protection.

Methods

The Independent TTP of the University Medicine Greifswald manages core functions: participant data administration, pseudonym management, consent and withdrawal administration and support for data requests for secondary research use. Recruitment involved representative sampling from population registries in all study regions. The TTP processed samples through multi-step validation, duplicate checks and pseudonym assignment before returning them to study centers for the invitation process. A detailed, modular informed consent was collected via a embedded iframe within the electronic data capture system used in the NAKO, signed by participants via biometric signpad, followed by the acquisition of confidentiality releases for treating physicians and social insurance data. The TTP manages pseudonym generation, administration, and resolution across all NAKO actors. Additionally, the TTP administers withdrawal requests, coordinates data deletion across institutions and facilitates secondary data access (insurance companies, tumor registry, statutory agencies).

Results

To date, the TTP supported the collection and management of approximately 600k modular consent declarations (comprising around 19 million individual consent items) and more than 580k releases from physicians' confidentiality. The TTP has also coordinated the withdrawal processes of more than 15k participants across all institutions holding NAKO data. In addition, it currently administers more than 16 million pseudonyms (up to 80 pseudonyms/participant) across the various NAKO processes and data flows.

Discussion

As a central component of the NAKO data infrastructure, the TTP is integrated into numerous data processing workflows, ensuring full DSGVO compliance. Currently the TTP integration via REST interfaces and web-form workflows streamlines the examination process by eliminating paper-based consent procedures and associated pre-examination verification steps. Deep system integration ensures that examination data cannot be captured without valid consent, automatically preventing protocol violations and strengthening both operational efficiency and regulatory compliance throughout the study.

[29-P1-08] Privacy Preserving Record Linkage for Outcome Enrichment in Real-World Data: A Registry-Mortality Use Case

Presenter: Alina Krichevsky

Authors: Matthias Drechsler [1]; Joceline Ziegler [1]; Alina Krichevsky [1]; Angela P. Moissl-Blanke [2]; Winfried März [3, 4, 5]; Anna Rubinski [1]

Affiliations: [1] Health Data Technologies GmbH ("Honic"), Neckarsulm; [2] Department of Medicine I, Medical Faculty Mannheim, University of Heidelberg, Mannheim; [3] Medical Clinic III, University of Heidelberg, Mannheim; [4] Department of Medicine, Medical University of Graz, Graz; [5] SYNLAB Academy, Synlab Holding Deutschland GmbH, Mannheim

Introduction

Secondary use of real-world data (RWD) is often limited by incomplete outcome information and fragmented data silos. In Germany, strict data protection regulations and restricted access to personal identifiers pose additional challenges for record linkage. Honic implements a privacy-preserving Master Patient Index (MPI) in collaboration with an external trust center, enabling deterministic linkage without direct identifier exchange.

This use case provides a unique opportunity to enrich outcome information in registry data through linkage with mortality dataset and to evaluate the quality of deterministic MPI-based linkage in comparison with probabilistic linkage approaches.

Methods

Data from the LURIC registry, a prospective cardiovascular cohort (N = 3,316 patients), were linked to a nationwide mortality dataset. Deterministic linkage relied on exact MPI matches, while probabilistic linkage applied fuzzy string matching and machine learning methods based on identifiers such as name, date of birth, sex, and address. A combined approach retained only matches consistent across both methods. Processing of personal identifiers in the registry data was permissible based on the consent given by the affected individuals. Use of identifiers in the mortality dataset was not restricted under the GDPR, as the GDPR does not apply to data of deceased persons (Recital 27). In the absence of a full ground truth, agreement between methods served as an indication of linkage quality. Manual validation of matches is ongoing.

Results

Deterministic MPI-based linkage identified 342 matched patients, corresponding to 10.3% of the registry population. These patients were considered deceased based on matching records in both data sources. Probabilistic linkage identified 379 patients (11.4%), and the combined approach identified 327 patients (9.9%).

Discussion

Privacy-preserving MPI-based deterministic linkage can effectively enrich outcome information in RWD while complying with data protection requirements. Linking registry data with mortality information addresses a frequently missing outcome, particularly in outpatient populations. Overall, this method enhances data completeness and quality for secondary use of RWD and offers a scalable framework for secure record linkage with potential applications in broader outpatient research settings.

[29-P1-09] Generalising Methods To Implement a Federated Trusted Third Party - the Merge'n'Dice Project

Presenter: Christopher Hampf

Authors: Christopher Hampf [1]; Martin Bialke [1]; Thomas Schult [1]; Ronny Schuldt [2]; Nico Wöller [1]; Marvin O. Kampf [3]; Philipp Heinrich [4]; Hauke Schneiderheinze [5]; Johnannes Wagner [5]; Holger Stenzhorn [6]; Jörg Römhild [7]; Torsten Leddig [1]; Dana Stahl [2]; Wolfgang Hoffmann [1]

Affiliations: [1] Institute for Community Medicine, Department Epidemiology of Health Care and Community Health, University Medicine Greifswald, Greifswald; [2] Trusted Third Party of the University Medicine Greifswald, Greifswald; [3] Medical Center for Information and Communication Technology, Universitätsklinikum Erlangen, Erlangen; [4] Trusted Third Party, Faculty of Medicine Carl Gustav Carus, TUD Dresden University of Technology, Dresden; [5] Trusted Third Party of the University Medical Center Hamburg-Eppendorf, Hamburg; [6] Data Integration Centre, Saarland University Medical Center, Homburg/Saar, Homburg; [7] University Hospital Tübingen, Tübingen

Introduction

Key to personalized medicine is knowledge of the complete patient journey. Therefore, the combination of data from different sources is paramount. A major challenge is linking data sets without a common identifier in a way that complies with data protection regulations. The Merge'n'Dice project [1] addressed this problem and was funded from 2023 to 2026 by the Federal Ministry of Research, Technology and Space as part of the 'Datentreuhandmodelle' funding programme (grant number 16DTM231). The aim was to generalise technical approaches of the federated Trusted Third Party (fTTP) to enable secure, cross-location data consolidation for various use cases, in compliance with data protection regulations.

Methods

A uniform HL7 FHIR interface for privacy-preserving record linkage (PPRL) and pseudonymisation was designed to implement different use cases. To identify the necessary use cases, a requirements analysis with stakeholders was carried out. A technical evaluation was conducted to validate the interface's practical suitability in a decentralised scenario with a focus on interoperability.

Results

After adding a 'pseudonymizePatient' interface to the 'TTP-FHIR Gateway' all use cases can now be handled uniformly: (1) centralised approaches using identifying data (e.g. in clinical trials) (2) decentralised approaches using coded data (e.g. MII and NUM-RDP) and (3) decentralised approaches using identifiers (e.g. the research data centre of the Federal Institute for Drugs and Medical Devices in Germany (BfArM)). The generic interface [2] was successfully tested with the support of five locations of the Medical Informatics Initiative (MII).

Discussion

The results demonstrate that the generalised fTTP interface can cover a wide range of use cases via a uniform interface. This reduces project-specific implementation efforts and ultimately enables faster location networking and merging of data (record linkage). Merge'n'Dice therefore makes a significant contribution to digital sovereignty in research, offering a model that transcends individual solutions. Future work can build on these generalised interfaces to further advance the networking of national research data infrastructures.

References

1. Merge'n'Dice - Unabhängige Treuhandstelle [Internet]. [cited 2026 Mar 12]. https://www.ths-greifswald.de/projekte/mergendice/. Accessed 12 Mar 2026

2. pseudonymizePatient - v2025.2.0 [Internet]. [cited 2026 Mar 12]. https://mosaic-hgw.github.io/ttp-fhir-ig-dispatcher/main/OperationDefinition-PseudonymizePatient.html. Accessed 12 Mar 2026

[29-P1-10] CancerLink - Use of PPRL approaches for data linkage in three selected cancer registries

Presenter: Christopher Hampf

Authors: Christopher Hampf [1]; Anna Krauß [1]; Alexandra Stein [1]; Ronny Schuldt [2]; Nico Wöller [1]; Sarah Theen [3]; Daniel Dettmer [4]; Lukas Mittnacht [5]; Martin Bialke [1]; Torsten Leddig [1]; Kerstin Weitmann [1]; Dana Stahl [2]; Wolfgang Hoffmann [1]

Affiliations: [1] Institute for Community Medicine, Department Epidemiology of Health Care and Community Health, University Medicine Greifswald, Greifswald; [2] Independent Trusted Third Party of the University Medicine Greifswald, Greifswald; [3] Cancer Registry Mecklenburg-Western Pomerania, Institute for Community Medicine, University Medicine Greifswald, Greifswald; [4] Clinical Cancer Registry of Lower Saxony, Hannover; [5] Cancer Registry Rhineland-Palatinate, Mainz

Introduction

Data from cancer registries are an important source of information for cancer research. Optimal use of the information collected by the registries, requires the option to request data from various cancer registries simultaneously. So far there is no systematic duplicate detection with subsequent record linkage. As a result, the data sets may contain unrecognized homonym errors or false-positives (entries are incorrectly assigned to the same patient, although they actually refer to different patients) and synonym errors or false-negatives (entries are incorrectly not merged, although they actually belong to the same patient). These errors can lead to problems in the analysis and thus significantly impair the quality and informative value of the data set.

Methods

For this project, data on selected cancer entities (ICD-10: C18, C19, C20, C34, C43, C53, C54, C55, C61) with a diagnosis in 2018-2024 were requested from three cancer registries: Mecklenburg-Western Pomerania, Rhineland-Palatinate, and Lower Saxony. By using a federated Trusted Third Party (fTTP) approach, it is possible to determine the number of duplicate individuals between cancer registries without sending of personally identifiable information (PII), resulting in a high level of data protection. By requesting a special RÜD variable that indicates whether cross-register data exchange has taken place, it is possible to determine how many of these duplicates were known to the cancer registries themselves.

To enable data merging, fTTP implements privacy-preserving record linkage (PPRL) based on Bloom filters. Using E-PIX® software, Bloom filters were generated locally from the PII within the participating cancer registries. The fTTP uses these Bloom filters to determine whether individuals are potential duplicates. For each identified individual the fTTP generates pseudonyms, which are reported back to the cancer registries. The cancer registries assign these pseudonyms to the requested registry data. In this way, it is possible to investigate whether there are clusters of homonym or synonym errors depending on parameters such as tumor entity, location, stage, gender, or age.

Results

The project is ongoing, but is expected to be finished in May 2026, so the results will be available for the conference.

Discussion

See the 'Results' section above.

[29-P1-11] Determining the date of death in the Bavarian population register for use at the University Hospital

Presenter: Timo Fuchs

Authors: Timo Fuchs [1, 2]; Thomas Ganslandt [3]; Jonathan M. Mang [1]

Affiliations: [1] Medical Center for Information and Communication Technology, Universitätsklinikum Erlangen, Erlangen; [2] TTZ Stein, Ansbach University of Applied Sciences, Ansbach; [3] Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen

Introduction

For many scientific questions in medicine, the survival rate of a patient cohort is highly relevant. To determine this, the patient's vital status date of death is required however, treating facilities only have access to hospital mortality data. In Germany, approximately 1 million people died in 2024, of whom about 430,000 died in a hospital. Consequently, the treating facility is aware of a patient's death in only 43% of all cases where a patient has already died. For this reason, cross-referencing survival data with the official population registry is essential for medical research at a university hospital.

Methods

To implement this, the legal framework was first examined, as specified in detail in the German Federal Registration Act (BMG). It distinguishes between basic registration information, which merely confirms that a person has died, and expanded registration information, which provides the specific date of death. To obtain expanded registration information, a legitimate interest must be proven. Since registration is a matter for the states, Bavarian laws were also taken into account.

Technical implementation is carried out via the API operated by the "Anstalt für kommunale Datenverarbeitung" (AKDB), as a service provider for access to the resident registry in Bavaria, using OSCI transport messages, which contain XMeld messages as payload with the corresponding DSMeld fields.

Results

An automated query of the population register is both legally and technically feasible. The connection has been tested as a prototype. The development of an automated query tool for this purpose is still pending.

Discussion

Cross-referencing the local patient cohort with the central population registry to determine patients' dates of death can make a valuable contribution to medical research. However, the productive implementation of this proof-of-concept is still pending. It could be made possible by developing a tool that other Bavarian university hospitals could use in a similar manner. Special cases (information blocks, address changes, multiple first and last names, and different spellings of names) and the volume of data, when querying all patients, have not yet been taken into account.

[29-P1-12] Evaluation of a Secure Pipeline for Automated DICOM Pseudonymization

Presenter: Alexa Iancu

Authors: Alexa Iancu [1]; Maximilian Schmidt [2]; Alexander Cavallaro [2]; Thomas Ganslandt [3]; Jonathan M. Mang [1]; Lorenz Kapsner [1]

Affiliations: [1] Uniklinikum Erlangen, Medical Center for Information and Communication Technology, Erlangen; [2] Institute of Radiology, Universitätsklinikum Erlangen, Erlangen; [3] Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen

Introduction

De-identification of Protected Health Information (PHI) in medical images is an important prerequisite for interinstitutional data exchange and research using medical images. In daily routine, workflows relying on manual procedures that are time-consuming and prone to user errors still exist.

Methods

To address this issue, an automated DICOM pseudonymization pipeline was developed to simplify and accelerate the preparation of imaging data for research and data sharing. The pipeline enables users to queue imaging studies by sending them directly to a dedicated DICOM node. Pseudonymization is then performed automatically using an already existing open-source pseudonymization framework with a pre-defined whitelist that has been approved by the hospital's data protection officer. The integration of the local Trusted Third Party (TTP) ensures secure identity management and enables linkage with secondary data. To evaluate the impact on the existing workflow, a user evaluation is currently being conducted among hospital staff who regularly perform pseudonymization tasks. The participants are completing questionnaires in which they compare the existing DICOM pseudonymization workflow with the newly developed automated pipeline in terms of general usability, time efficiency, error risk and overall satisfaction.

Results

The automated pipeline decreases the workload for users allowing imaging studies to be submitted directly to the designated DICOM node. By reducing the pseudonymization process to one single step, this approach eliminates multiple manual steps, including data export and parameter configuration. Hence, the developed pipeline can easiliy be integrated into clinical practice and generates pseudonymized DICOM metadata that can be utilized directly for a wide variety of research purposes. Initial user feedback is currently being collected to assess usability, efficiency, perceived error risk and general acceptance in clinical practice.

Discussion

The presented approach demonstrates how existing pseudonymization tools can be integrated into existing worflows in radiology to reduce manual effort and complexity. While the formal results of the user survey are still pending at the time of submitting this abstract, the technical implementation already highlights the potential to improve efficiency and reduce errors in data preparation for research projects. Future work will include a comprehensive assessment of the evaluation questionnaire among users, aiming for a broader clinical adoption.

[29-P1-13] Metabase Dashboard for Workload Management in Data Integration Centers Based on Data Request Complexity

Presenter: Yevgeniia Ignatenko

Authors: Yevgeniia Ignatenko [1]; Natalia Ortmann [2]; Dorian Quell [1]; Cise Nil Köksal [1]; Sinah Heinz [1]; Ludwig Christian Hinske [1]; Iñaki Soto Rey [2]

Affiliations: [1] Digital Medicine, University Hospital of Augsburg, Augsburg; [2] Digital Medicine, University Hospital of Augsburg, Augsburg, Germany., Augsburg

Introduction

Medical Data Integration Centers (DICs) face increasing numbers and growing complexity of data requests for research projects. This complicates workload planning and can lead to uneven task distribution within teams. In project management contexts, this is associated with risks such as resource bottlenecks, project delays and reduced transparency.

Existing project management systems typically do not properly account for project complexity and provide limited support for effective workload management. This work aims to develop a dashboard for transparent workload management based on data request complexity, including early identification of potential risks.

Methods

Dashboard was designed and developed through internal workshops to ensure usability and integration into existing workflows. These workshops were used to define user requirements, key metrics, and visualizations.

A central component is a complexity scoring system for data requests, created iteratively. The score considers criteria such as project type, number of variables, involved data sources, standardization, required expertise, data availability, and output complexity. Based on this score, data requests are categorized into three levels representing estimated effort (up to 8, up to 40, more than 40 hours).

Dashboard was implemented using Metabase, a web-based BI tool and integrates data from the Taiga project management system. It visualizes workload per team member, data request complexity, and request status, with filtering options across all data.

Results

Dashboard integrating data requests and team member assignments was developed to support workload distribution and planning. It provides an overview of incoming data requests within sprints, their processing status, and grouping by request type or department. It also includes metrics such as processing times per step and average processing times across projects.

Dashboard improved transparency of workload distribution among team members through visualization. It may enhance planning, promote a more balanced workload, and enable early identification of risks, such as the concentration of high-complexity requests among individual team members.

Discussion

Integration of the dashboard into existing workflows supports job-flow and sprint planning. By incorporating data request complexity, approach may support more informed, data-driven decision-making. Future work will evaluate its impact on workload distribution and balance across team members.

[29-P1-14] Privacy-Compliant Integration of Clinical Routine Data in Interdisciplinary Research Projects

Presenter: Yevgeniia Ignatenko

Authors: Natalia Ortmann [1]; Johannes Raffler [2]; Andreas Mahler [3]; Kathrin Ebertsch [3]; Roman Hutfluss [4]; Emily Schiller [5]; Iñaki Soto-Rey [6]; Yevgeniia Ignatenko [7]

Affiliations: [1] Augsburg University Hospital, 86156 Augsburg; [2] Institute for Digital Medicine, University Hospital Augsburg, Germany, Augsburg; [3] Department of Digitalization and Nursing Science, University Hospital of Augsburg, Augsburg; [4] Digital Health & Medical Decision Making, University of Augsburg, Faculty of Business and Economics, Augsburg; [5] XITASO GmbH IT & Software Solutions, Augsburg, Augsburg; [6] Digital Medicine, University Hospital of Augsburg, Augsburg; [7] Augsburg University Hospital, Augsburg

Introduction

Clinical routine data hold significant potential for research and data-driven decision support but are often constrained by technical and regulatory barriers [1]. Their secondary use requires robust governance structures, traceable data processing, and secure infrastructures, particularly when involving interdisciplinary collaborations between clinical practice, academia, and industry [3].

This article presents a practical GDPR-compliant approach to enabling the use of clinical routine data in interdisciplinary research projects. The approach was successfully piloted in the context of the KISPP project (AI-driven system for demand-based nursing staff planning and reporting).

Methods

Under the direction of the Department of Digitalization and Nursing Science at the University Hospital Augsburg (UKA), a core dataset comprising 42 structured variables was defined. The dataset was subsequently validated by the Medical Data Integration Center at UKA (MeDIZ). To ensure GDPR-compliant use of sensitive health data and legally compliant cooperation between all involved parties, a comprehensive data protection and data usage concept was developed in collaboration with all relevant stakeholders in accordance with established recommendations for medical research projects [2].

Results

The workflow starts with data extraction and integration by MeDIZ, followed by de facto anonymization of the dataset. The data were then made available within a secure research infrastructure (high-performance computing cluster) at the Institute for Digital Medicine. External project partners received temporary access to this controlled environment, enabling data analysis and AI model training without transferring patient level data outside the hospital infrastructure. All data access and processing activities were monitored and logged, ensuring accountability and traceability of data use [3].

Discussion

The results show that using clinical routine data in interdisciplinary research projects is feasible but associated with regulatory and organizational challenges. In particular, approval processes related to data protection and legal requirements were time-consuming and required early involvement of all relevant stakeholders. Nevertheless, a robust data protection and data usage concept was successfully established. Combined with the secure research environment at UKA, this enabled controlled data access for external academic and non-academic partners for AI-based research purposes. The proposed framework may serve as a transferable and scalable model for future interdisciplinary research projects involving internal and external partners.

[29-P1-15] Current Text Anonymization Approaches Are Not Good Enough: Limitations in Medical Text Processing

Presenter: Roland Roller

Authors: Roland Roller [1]

Affiliations: [1] Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Berlin

Introduction

Anonymization of clinical text is essential for data sharing and secondary use in medical informatics and research. Many approaches conflate anonymization with de-identification, removing explicit identifiers while leaving indirect identifiers and quasi-identifiers unprotected. Detection of direct and indirect identifiers often relies on named entity recognition (NER) and is never 100% accurate due to language nuances, creating potential privacy risks. Stylometric patterns and contextual details further increase re-identification risk.

Methods

We conduct a qualitative analysis on 63 fictitious discharge summaries (GraSSCO) with common anonymization strategies, including detecting direct and indirect identifiers, LLM-based text rewriting, and differential privacy (DP)-based approaches (DPMLM).

Results

Results show that the detection of direct or indirect identifiers helps remove or generalize sensitive information but relies on imperfect models. In our case, we achieve only an F1 score of 0.96 for direct and 0.79 for indirect identifiers, missing potentially critical information - though not all missed identifiers are equally important. Text rewriting with LLMs may reduce indirect risks but does not guarantee privacy, and medically relevant information can be altered, affecting data utility. DP can provide formal guarantees however, at levels sufficient for safety, it often destroys readability and relevant medical content, making outputs clinically unusable. In the following an example with secure DP-modification (epsilon=10):

(org) The rehabilitation treatment was significantly complicated, as Ms. Mitchell, due to advanced dementia, was unfortunately no longer able to reliably perform even simple movement sequences independently, even with intensive and repeated training.

(DP) Aud dream apparatus was ' fun, as Mei David, leve to subsequent decline, was seen no panel unjust to ep opoly archive between caption tracts iots, don with landsc and pearson latino.

Discussion

All current approaches face a trade-off between privacy and utility. Detection-based techniques are imperfect, and text rewriting cannot guarantee privacy. DP ensures safety but may render medical content unusable. Text anonymization should focus in future on controllable, targeted DP approaches that selectively modify sensitive information while preserving the remainder, balancing privacy with clinical interpretability. Moreover, we believe that standardized evaluation frameworks are needed to jointly assess privacy, residual risk, and utility.

[29-P1-16] A Framework to Evaluate Synthetically Generated Data

Presenter: Paolo Messina

Authors: Alessandro Cesa [1]; Paolo Messina [1]; Mario Torchia [1]; Luca Emili [1]

Affiliations: [1] InSilicoTrials Technologies Spa, Trieste

Introduction

Evaluating synthetically generated data is a critical step in ensuring that synthetic data generation methods produce outputs that are trustworthy, useful, and safe. Unlike supervised learning, where model performance can be directly assessed against ground-truth labels, synthetic data evaluation does not rely on a single reference target. Instead, quality must be assessed across multiple dimensions, each capturing different requirements.

Methods

We defined a unified and extensible framework for the evaluation of synthetically generated tabular datasets. In line with the prevailing literature, our framework is structured around three key evaluation pillars: privacy, fidelity, and utility.

Privacy assesses whether the synthetic data sufficiently protects the individuals represented in the original (training) dataset. Privacy was quantified through three complementary risks. Singling out risk measures the possibility of identifying a rare combination of attributes in the synthetic data that corresponds to a real observation. Linkability risk that evaluates whether records from the synthetic dataset can be linked to records in the original dataset through shared quasi-identifiers. Inference risk, that captures the extent to which sensitive attributes about individuals in the original data can be inferred from the synthetic data.

Fidelity measures how closely the statistical properties of the synthetic data match those of the original data, including distributions, correlations, and higher-order relationships.

Utility evaluates whether downstream tasks performed using synthetic data achieve performance comparable to, or better than, those performed using the original data.

All metrics are mapped to normalized scores, enabling consistent comparison and aggregation across evaluation dimensions.

Results

The framework was implemented as a scalable software suite and deployed as a reusable pipeline component, enabling automated, reproducible, and efficient evaluation within data-generation workflows. To improve scalability and efficiency, the framework leverages multiprocessing to compute evaluation metrics in parallel, significantly reducing execution.

Discussion

This framework provides a standardized, modular and extensible, approach to a complex multi-dimensional evaluation task, enabling the end user to easily evaluate synthetically generated data, across the three key evaluation pillars: privacy, fidelity, utility. Its pipeline-ready design makes it broadly applicable across domains where trustworthy synthetic data generation is essential.

[29-P1-17] Towards a Secure Processing Environment - Results from a Scoping Review in the KIMed Project

Presenter: Maryam Khodaei Dolouei

Authors: Maryam Khodaei Dolouei [1]; Ronny Schubert [2]; Frank A. Meineke [1]; Stefan Bohn [3]

Affiliations: [1] Institut für Medizinische Informatik, Statistik und Epidemiologie, Universität Leipzig, Leipzig; [2] Saxon Institute for Computational Intelligence and Machine Learning (SICIM), University of Applied Sciences Mittweida, Mittweida; [3] Innovation Center Computer Assisted Surgery, Faculty of Medicine, Leipzig University, Leipzig

Introduction

The European Health Data Space (EHDS) establishes a new regulatory framework for the secondary use of health data, with Secure Processing Environments (SPEs) at its core. However, secure data processing for research is not a novel concept - Trusted Research Environments (TREs) and similar models have already been implemented across the research community. These approaches vary significantly in terminology, technical design, and governance. Although the EHDS provides a specific definition and regulatory framework for SPEs, it remains unclear how existing implementations align with this new standard, which elements can be reused, and where key differences lie. As part of the collaborative KIMed - Network for Artificial Intelligence in Medicine project, we are conducting a scoping review to lay the foundation for an EHDS-compliant SPE in Saxony, Germany.

Methods

Following the Joanna Briggs Institute and PRISMA-ScR guidelines, we conducted a comprehensive search across Web of Science, PubMed, IEEE Xplore, and Scopus (January 2010 - December 2025).

The review includes both implemented SPEs and TREs, as well as conceptual, prospective, and evaluative studies on their design and use. We analyze technical architectures, user and provider requirements, governance mechanisms, and the integration of machine learning. Data extraction is currently underway, extracting over 50 key characteristics per article, including alignment with the 5-Safes framework.

Results

A total of 183 peer reviewed articles were identified, 25 of which were considered relevant to our research question. Preliminary screening reveals 23 studies on European SPEs or comparable environments, including 7 with explicit links to the EHDS. Additional findings from North American implementations will provide valuable comparative insights. The ongoing data extraction phase is systematically cataloging technical and organizational dimensions of SPEs.

Discussion

Due to the recent emergence of the EHDS, literature on its implementation remains limited. Diverse national approaches, varying data formats, and differing security requirements highlight the need for networking to share insights and learn from early adopters. This scoping review identifies trends and key challenges and aims to lay the foundation for the development of an EHDS-compliant SPE for Saxony.

[29-P1-18] Secure Processing Environments as a Bridge Between Data Protection and International Health Research using INTER.SECT as an example

Presenter: Karla Riesterer

Authors: Karla Riesterer [1]; Kayvan Bozorgmehr [2]; Johannes Drepper [3]; Marie-Christine Laible [4]; Bianka Detering [2]; Moritz Steiner [1]; Jan Oppenberg [2]; Eilin Rast [2]

Affiliations: [1] TMF - Technologie- und Methodenplattform für die vernetzte medizinische Forschung e.V., Berlin; [2] Fakultät für Gesundheitswissenschaften, AG 2 Bevölkerungsmedizin und Versorgungsforschung, Universität Bielefeld, Bielefeld; [3] TMF - Technologie- und Methodenplattform für die vernetzte medizinische Forschung e. V., Berlin; [4] Bundesamt für Migration und Flüchtlinge, Nürnberg

Introduction

Secure Processing Environments (SPEs) are gaining importance amid growing regulatory requirements, as they offer a technically safeguarded setting for research using sensitive health data. Several German authorities provide access to selected administrative data within SPEs, including the Federal Office for Migration and Refugees (BAMF). These controlled environments allow researchers to analyse data without physically transferring datasets, supporting secure processing while protecting individuals' rights and privacy.

The INTER.SECT project examines how pre-migration health status and post-migration contextual factors shape the health and mortality of resettled refugees in Germany. Because this work requires

the processing, anonymisation, and potential linkage of sensitive health data with administrative BAMF records, all analyses are conducted within a SPE. This makes a comprehensive data protection assessment essential.

This abstract describes the procedure and outcome of a voted data protection concept developed for INTER.SECT and discusses the role of SPEs in enabling such research in a legally compliant and secure manner.

Methods

A detailed data protection concept, outlining three possible implementation scenarios was developed through an iterative, interdisciplinary process in consultation with a designated rapporteur and submitted to the Working Group on Data Protection (WG-DP) at TMF e.V. for review and vote.

The evaluation focused on the technical and organisational measures within BAMF's SPE, the methodological safeguards for anonymisation, and the coherence of the risk analysis. The legal admissibility of BAMF's specific legal bases and internal IT security architecture remained outside the scope.

Results

The WG-DP voted positively and concluded that the concept is coherent, comprehensible, and methodologically sound, and that no data protection concerns prevent the project from proceeding, provided considerations for further consultation and implementation for internal BAMF decisions are addressed.

Regardless of the implemented scenario, BAMF's SPE features enhanced technical and organisational safeguards as well as a dedicated physical secure facility, thereby enabling controlled research access, allowing even highly sensitive administrative data to be used for scientific research.

Discussion

The procedure highlights challenges of linkage-based public-interest research, including complex legal frameworks, limited disclosure of IT-security infrastructure, and scenario-dependent risks. It also indicates that SPEs, embedded in robust governance and independent expert oversight, can enable sensitive administrative health research.

[29-P1-19] Implementation of a Centralized Objection Office for Patient Data Use in Research

Presenter: Dana Stahl

Authors: Dana Stahl [1]; Martin Bialke [2]; Torsten Leddig [2]; Wolfgang Hoffmann [3]

Affiliations: [1] Universitätsmedizin Greifswald, Greifswald; [2] Institute for Community Medicine, University Medicine Greifswald, Greifswald; [3] Universitätsmedizin Greifswald KöR, Institut für Community Medicine, Greifswald

Introduction

The implementation of a centralized objection office for patient data use in research has provided valuable insights into the integration of legal, technical, and organizational requirements in health research data governance.

Methods

The experience at the University Medicine Greifswald highlights that compliance with data protection laws, such as §50 LKHG MV, requires more than legal understanding-it demands technical integration, clear organizational structures, and sustainable processes.

Results

The interim solution currently in place enables basic compliance but is not scalable and increases long-term compliance risks, emphasizing the need for proactive planning and investment in robust infrastructure. Collaboration with existing data governance structures, such as the Trusted Third Party and the Data Integration Center, is crucial, as they offer proven expertise, infrastructure, and processes that can accelerate implementation and reduce costs.

Discussion

However, such collaboration must be formalized to ensure clear roles and responsibilities. The UMG's experience in implementing an opt-out process for research with clinical health data provides a solid foundation for establishing similar mechanisms in the European Health Data Space (EHDS). This allows other institutions or countries connected to the EHDS to learn from the UMG's approach.

The full paper pertaining to this abstract will be published in the GMDS journal "Medizinische Informatik, Biometrie und Epidemiologie (MIBE)."

[29-P1-20] Digital consent process for minors: expansion of gICS® to enable flexible capture of multiple signatures

Presenter: Dana Stahl

Authors: Kathleen Dittmann [1]; Arne Blumentritt [1]; Martin Bialke [2]; Frank-Michael Moser [1]; Lizon Fiedler-Lacombe [1]; Ekatarina Heim [1]; Dana Stahl [1]

Affiliations: [1] Universitätsmedizin Greifswald, Unabhängige Treuhandstelle Greifswald, Greifswald; [2] Institute for Community Medicine, University Medicine Greifswald, Greifswald

Introduction

The participation of minors in clinical trials presents unique legal and organizational challenges. Of particular importance is the legally binding consent from legal guardians. For over 10 years, Unabhängige Treuhandstelle Greifswald (THS) has offered generic Informed Consent Service (gICS®), a digital solution for managing informed consent forms. Due to growing importance of research involving vulnerable groups, particularly minors, the technical capabilities of gICS® must be adapted

to their specific requirements - especially regarding to collection of multiple signatures and age appropriate form design.

Methods

As smallest unit for defining scope of consent, such as data use or sample collection, gICS® uses so-called "policies". Until now, validity of these elements was configured based on study duration rather than age of participants. In addition, signatures by legal guardians are mandatory for minors. It was necessary to expand the system´s functionality to allow the recording of different signatory roles (e.g. two legal guardians), age-related signature requirements, and expiration dates upon reaching age of majority. Furthermore, it should be possible to evaluate these automatically using rule-based mechanisms.

Results

The system has been expanded to include group-specific and type-oriented configuration logic. This allows several legal guardians of minor to sign the same template. For example, for six-year-old children, only an information disclosure form is provided, without a requirement to sign. The date of reaching age of majority is calculated dynamically and based on rules, using participant's date of birth. The expiration date is calculated automatically, and the date of birth itself is not stored for reasons of data minimization. Both, new signature concept, and new rule-based mechanisms, enable precise and legally compliant handling of consent in studies involving minors.

Discussion

The expansion of gICS® enhances its usability for complex consent processes, even when dealing with vulnerable groups. The flexible design of signature fields can be easily adapted to other scenarios, such as legal representatives of patients who are unable to give consent. With its digital solution, which complies with § 126b BGB, gICS® helps improve quality and transparency of consent practices in medical research. As a free solution, gICS® is available to researchers nationwide (https://www.ths-greifswald.de/gics).

[29-P1-21] BC-Minors: An Automated Workflow for Re-Contacting Former Minor Patients Under Broad Consent Once They Reach The Age Of Majority

Presenter: Daniel Schröder

Authors: Daniel Schröder [1]; Seyedeh Delaram Mirraziroudsari [1]; Tobias Brix [1]; Sascha Prus [2]; Michael Storck [3]; Dominik Heider [1]

Affiliations: [1] Institute of Medical Informatics, University of Münster, Münster; [2] University of Münster, University Hospital Münster, Münster; [3] Universität Münster, Münster

Introduction

The German Medical Informatics Initiative introduced the Broad Consent (BC) as a standardized framework for the secondary use of patient data and biomaterials in medical research [1]. Applying BC to minors entails additional legal and organizational requirements, as individuals whose consent

was originally given by their legal guardians must be informed of their right to withdraw that consent once they reach the age of majority. In large hospitals, this process is labor-intensive and error-prone when performed manually. To address this challenge, we developed BC-Minors, an automated workflow for identifying and informing former minor patients.

Methods

We analyzed the regulatory requirements for re-contacting former minor patients and reviewed the local technical and organizational infrastructure available for this task. Based on this, we defined a standardized workflow that covers data sources, timing of re-contact, the required content of the letters, and responsibilities for drafting, printing and mailing.

Results

BC-Minors was implemented as a Python-based script executed at regular intervals. It queries the consent database and hospital information system (HIS) to identify patients who signed a minors' BC and are approaching their 18th birthday. For these patients, the system retrieves the most recent postal address and the originally approved consent modules. Based on predefined templates, personalized PDF letters are automatically generated, containing a blank copy of the signed BC document and a QR code linking to the BC withdrawal webpage. Files are stored on a secure network share, forwarded to the institutional printing service, and distributed via the hospital mailroom. All processing steps are logged.

Discussion

Practical challenges remain. Postal addresses may be outdated or no longer valid, especially for patients who have not been to the hospital in years. In addition, young people who sign the form shortly before their 18th birthday may receive a letter shortly thereafter asking them to contact the hospital again. Postal communication also causes recurring printing and mailing costs of approximately three euros per letter. Nevertheless, BC-Minors demonstrates a feasible technical approach for operationalizing legal requirements related to minors' BC transitions.

References

[1] Semler SC, Wissing F, Heyder R. German Medical Informatics Initiative. Methods Inf Med. 201857(Suppl 1):e50-e56. doi:10.3414/ME18-03-0003.

[29-P1-22] Privacy-Preserving Data Sharing in OSSE-Based Registries: A Prototype Extension for Anonymization and Federated Search

Presenter: Jessica Vasseur

Authors: Axel Zieschank [1]; Jessica Vasseur [1]; Jens Göbel [1]; Holger Storf [1]

Affiliations: [1] Goethe-Universität Frankfurt, Universitätsmedizin, Institut für Medizininformatik, Frankfurt am Main

Introduction

Health data are among the most sensitive personal data categories, carrying an elevated re-identification risk. This also applies to data stored in medical registries such as those based on the Open Source Registry System for Rare Diseases (OSSE) [1]. Anonymization represents a critical strategy for protecting individual privacy and preventing data misuse. As part of the PrivacyUmbrella project (funded by BMFTR, 16KISA003), we developed a prototypic extension of OSSE to allow anonymized data provision from OSSE-based registries.

Methods

As part of the project OSSE was extended in several steps. First, the established REST-based API of OSSE was extended with a new machine-readable tabular export interface, complementing the existing human-readable export functionality. Secondly, an interface was developed to enable in-registry anonymization and direct retrieval of anonymized data, with anonymization performed by the ARX software library [2]. This was complemented by a heuristic approach that accounts for structural anonymity in graph-like data. Thirdly, the search broker for distributed search was redesigned to integrate with the REST-based API of OSSE.

Results

The resulting demonstrator comprises prototypic OSSE-based registry instances, which provide anonymization interfaces that account for the structural complexity of registry data. The tabular REST-based export improves machine processability and facilitates integration with established anonymization tools such as ARX. Users can configure data element definitions - specifying, for instance, which fields to include, and whether fields represent (pseudo-)identifiers - as well as the anonymization strategy. To support cross-registry use, a federated search with a central search broker was prototypically implemented, enabling registry operators to retrieve pending queries and respond with anonymized data in a controlled manner while preserving data sovereignty.

Discussion

The developed extensions of OSSE lower the barrier to anonymizing data from OSSE-based registries and provide a foundation for privacy-preserving federated data sharing, combining optimized anonymization strategies with controlled, distributed data access across registry boundaries.

References

[1] Storf H et al. Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz, 01 May 201760(5):523-531

[2] Prasser, Fabian, et al. 'Flexible data anonymization using ARX - Current status and challenges ahead.' Software: Practice and Experience 50.7 (2020): 1277-1304.

[29-P1-23] Sustainable archiving of rare disease registries: A conceptual framework based on the NEOCYST registry

Presenter: Selina Wilhelmi

Authors: Selina Wilhelmi [1]; Jens Göbel [2]; Holger Storf [2]; Jessica Vasseur [2]

Affiliations: [1] Goethe University Frankfurt, University Medicine, Institute of Medical Informatics (IMI), Frankfurt; [2] Goethe University Frankfurt, University Medicine, Institute of Medical Informatics (IMI), Frankfurt am Main

Introduction

Medical registries, such as the NEOCYST (Network for Early Onset Cystic Kidney Diseases) registry (1), generate highly valuable data but are often tied to limited project funding. After project completion, continued operation, including controlled data availability, is often not guaranteed. This raises the need for archiving strategies supporting sustainable data preservation, governance, and future reuse, also considering regulatory requirements such as the 10-year data preservation mandated by GCP for clinical trials.

Potential approaches include complete archiving (non-operational data and system preservation) and usable archiving (annotated data in a controlled, read-only environment for secondary use). Additionally, long-term data stewardship must align with FAIR principles while addressing ethical and organizational responsibilities. Here, we present a conceptual framework for archiving registry data, using the NEOCYST registry as use case (funded by BMFTR 2016-2026, 01GM2203F).

Methods

The analysis considered key components of registry systems based on the OSSE (Open-Source System for Rare Diseases) software (2), including identifying data, medical data, metadata, database structures, and governance aspects. Archiving strategies were systematically compared regarding data protection, long-term interpretability, reusability, implementation complexity, and FAIR compliance.

Results

Complete archiving, e.g., using SIARD (Software Independent Archival of Relational Databases), supports long-term preservation and future system reconstruction but limits data availability for research without additional expertise. Usable archiving ensures data accessibility but incurs additional costs and requires governance. Hybrid approaches balance preservation and reusability but increase system complexity.

Based on these findings, a modular framework is proposed comprising: 1) medical data, 2) standardized metadata, 3) relational database structure, and 4) contextual and governance documentation, including consent procedures, data protection, handling of personal identifying data, user roles and access restrictions.

Discussion

In rare disease registries, small cohort sizes increase re-identification risk, imposing additional constraints. Therefore, archiving must balance accessibility, data protection, and operational feasibility. Sustainable archiving requires not only technical solutions but also clearly defined responsibilities for long-term storage. The proposed framework supports structured decision-making and highlights the need for standardized approaches to post-project data handling in medical registries.

References

(1) König JC et al. Front. Pediatr., 13 February 20186:24

(2) Storf H et al. Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz, 01 May 201760(5):523-531

[29-P1-24] The use of the Polish National Health Fund (Silesian Voivodeship Branch in Katowice) registry data in medical research

Presenter: Michał Skrzypek

Authors: Michał Skrzypek [1]; Dorota Stołtny [2]; Alicja Nowowiejska-Wiewióra [3]; Daniel Cieśla [4]; Krzysztof Strojek [2]; Mariusz Gąsior [5]

Affiliations: [1] Department of Biostatistics, Faculty of Public Health in Bytom, Medical University of Silesia in Katowice (Poland), Bytom; [2] Department of Internal Medicine, Diabetology and Cardiometabolic Diseases, Silesian Centre for Heart Diseases, Zabrze; [3] 3rd Department of Cardiology, Silesian Centre for Heart Diseases, Zabrze; [4] Department of Science and New Technologies, Silesian Center for Heart Diseases, Zabrze; [5] 3rd Department of Cardiology, School of Medical Sciences in Zabrze, Medical University of Silesia, Zabrze

Introduction

The Polish National Health Fund (NHF) registry was originally developed to support financial management within the public healthcare system. However, its general coverage of medical events creates an opportunity for epidemiological and clinical research. As an example, we present an analysis of de novo type II diabetes mellitus (DM) among patients hospitalized for heart failure (HF) in the Silesia Voivodeship.

Methods

The NHF is the sole public payer for health services in Poland, and all providers are required to report medical events - including primary care and outpatient visits, hospitalizations, diagnoses (ICD-10), and procedures (ICD-9) - to obtain reimbursement. Based on an agreement between the Regional Silesian NHF Department in Katowice and the Centre for Heart Diseases in Zabrze, the SILCARD database was created. It includes records of all patients hospitalized for cardiovascular diseases or hospitalized in cardiology, cardiac surgery, vascular surgery, or diabetology units for any reason in the Silesia Voivodeship since 2006.

Data are pseudonymized, retaining only gender, date of birth, and coded residence. This allows longitudinal tracking of ICD-coded health events without identifying individuals. SILCARD is updated annually.

Results

The SILCARD database contains 1,083,481 patients. We identified adult patients with a first-time HF hospitalization (ICD-10 code I50) using a 5-year look-back period to exclude prior HF diagnoses and required at least 1 year of follow-up. A total of 171,880 patients hospitalized between 2011 and 2023

met these criteria. Among them, 38,889 (31.9%) had pre-existing type II DM (E11), 68,032 (55.8%) had no DM, and 15,085 (12.3%) developed de novo DM during or after the index hospitalization. Detailed characteristics and outcomes will be presented.

Discussion

Our study demonstrates that administrative health data collected for financial purposes can support large-scale medical research. Although NHF data provide near-complete population coverage, they lack clinical granularity compared with electronic health records (EHR). Additionally, access to these sensitive data remains highly restricted, which poses challenges for wider research applications.

[29-P1-25] Developing Computable Phenotypes for Pancreatic Cancer

Presenter: Tammo Klaaßen

Authors: Tammo Klaaßen [1]; Sophia Rheinländer [1]; Nils Beyer [1]; Christoph Ammer-Herrmenau [2]; Harald Kusch [1]; Ulrich Sax [1]

Affiliations: [1] University Medical Center Göttingen, Department of Medical Informatics, Göttingen; [2] University Medical Center Göttingen, Clinic for Gastroenterology, Gastrointestinal Oncology and Endocrinology, Göttingen

Introduction

Pancreatic cancer is one of the most lethal malignancies. According to the Robert Koch Institute, the 5-year survival rate in Germany was approximately 11% in 2023 [1]. During the clinical management of patients with pancreatic cancer, large amounts of health-related data are generated, offering considerable potential for clinical and translational research. However, these data are often not directly suitable for the identification of specific patient cohorts and may first require the derivation of computable variables from heterogeneous source data, as similarly described for diabetes research [2, 3]. One possible approach is the use of computable phenotypes, in which clinical characteristics are translated into formally defined, machine-processable rules.

Methods

Data from multiple sources were used for the implementation, including the Medical Data Integration Center of the University Medical Center Göttingen (UMG-MeDIC) and the Molecular Pancreatic Carcinoma Program (MolPAC). Clinically relevant phenotypes were identified on the basis of relevant clinical guidelines and expert knowledge provided by clinicians and were assessed regarding their feasibility and availability in the respective data sources. The selected phenotypes were subsequently translated into formalized rule-based definitions and operationalized as computable phenotypes. Finally, a prototype implemen- tation was developed to identify patient cohorts based on these computable phenotypes and to support future application to newly documented patients.

Results

Several clinically relevant computable phenotypes for pancreatic cancer were prototypically developed and implemented. Both clinical data and routine health care data were considered. In particular, computable phenotypes that enabled the formal representation and integration of heterogeneous data sources within a common analytical framework proved especially suitable.

Discussion

The results suggest that computable phenotyping is a promising approach for identifying clinically relevant patient groups from existing health data. How- ever, its application in pancreatic cancer has so far only been described to a limited extent and has not yet been standardized consistently. At the same time, the findings underline that the heterogeneity of the available data sources constitutes a major challenge for the development and implementation of robust computable phenotypes. References [1] Koch-Institut, 2025. doi: 10.25646/13129. [2] L. Reitzle et al., 2023, doi: 10.1055/a-1791-0918. [3] N. J. Thomas et al., 2023, doi: 10.1016/j.jclinepi.2022.10.022.

[29-P1-26] Analysis of electronic order sets for medication orders in a German university hospital

Presenter: Luise Kauk

Authors: Luise Kauk [1]; Hanna M. Seidling [1]; Janina A. Bittmann [1]

Affiliations: [1] Heidelberg University, Medical Faculty Heidelberg / Heidelberg University Hospital, Internal Medicine IX - Department of Clinical Pharmacology and Pharmacoepidemiology, Heidelberg

Introduction

The availability of electronic order sets is considered a key factor for the successful implementation of computerized physician order entry systems (CPOE). As a predefined set of orders to be prescribed simultaneously, they can standardize ordering processes and reduce workload. Order sets can contain several types of connected orders, e.g., for procedures, examinations, or medication and thus display an entire standardized treatment process. Although the usage of order sets is well documented internationally, little data is available on their structure and usage within German hospitals. This project aimed to analyse the structure and quality of existing electronic order sets for medication orders at a German university hospital and to identify areas for improvement.

Methods

All electronic order sets available in the implemented CPOE for general wards at Heidelberg University Hospital were descriptively analysed and classified into indication specific order sets (related orders for a specific diagnosis or treatment that can be ordered together) and medication lists (lists of similar orders for an indication, among which one or few orders are chosen). Using 19 literature-based quality criteria for order sets, options for improvement were identified.

Results

Of the overall 329 order sets, 54.7 % (n = 180) were classified as indication specific, containing an average of 4.9 $\pm$ 3.9 orders per order set and 45.3 % (n = 149) as medication lists with an average of 11.1 $\pm$ 10.1 orders per order set. Identified options for quality improvement were a reduction of the number of orders per order set to provide an easier overview of the content of each order set, changing the selectability from single orders within an order set to the entire set where possible and pre-filling all data required for a complete order to standardize and fasten the ordering process.

Discussion

The findings of this analysis can be used to improve the usability of the existing order sets for users, which could streamline the prescribing process and facilitate the implementation and usage of CPOE. For further improvement and to see how beneficial order sets actually are, their utilization rate should be made available through the system.

[29-P1-27] Calculating Contraindications from Health Care FHIR Data in the INTERPOLAR Project

Presenter: Alexander Strübing

Authors: Alexander Strübing [1]; Florian Schmidt [1]; Sebastian Stäubert [1]; Matthias Reusche [1]; Marcel von Borzestowski [2]; Tatjana Beppler [3]; Helene Köster [4]; Markus Löffler [1]; Daniel Neumann [1]

Affiliations: [1] Institute for Medical Informatics, Statistics and Epidemiology (IMISE), Leipzig University, Leipzig; [2] Universität Leipzig, Medizinische Fakultät, Leipzig; [3] Jena University Hospital, Jena; [4] Universitätsklinikum Erlangen, Erlangen

Introduction

Within Germany's Medical Informatics Initiative (MII), Data Integration Centers have been set up to enable researchers to access data through FHIR endpoints using the MII core dataset (CDS).

The MII project INTERPOLAR focuses on detecting medication-related problems (MRPs) and conducting a retrospective analysis of pharmacists' clinical decisions. Therefore the CDS toolchain has been established. Performing analyses, calculations and other data-related tasks there is a 'dataprocessor' module which operates on FHIR data which is available through a relational database. This article focuses on how contraindications are calculated from continuous patient data.

Methods

The data processor is organised into sub-modules, enabling independent analysis. The calculation of contraindication is done in submodule "MRP-Calculation". It consumes patient data, e.g. medication (ATC-coding), conditions (ICD10-coding), procedures (OPS-coding) and laboratory values (LOINC-coding) and is dependent on the temporal context, which is determined by a medication analysis conducted by the station's pharmacist. These data are then algorithmically matched with lists consisting of pairs of codes for drugs that interact negatively with each other (Drug/Drug, Drug/Drug-Group), drugs and diagnoses (Drug/Disease) as well as drugs and laboratory values (Drug/Kidney).

The challenge is to convert units appropriately, consider patient- and laboratory parameter-specific reference ranges, and use proxy indicators for diagnoses, such as laboratory values or procedures. The found contraindications are stored together with their provenance information and are then available for validation by pharmacists.

Results

The submodule entitled 'MRP-Calculation' is available to the public on GitHub [(https://github.com/medizininform initiative/INTERPOLAR/tree/main/R-dataprocessor/submodules/02_MRP_Calculation). The application is currently being utilised within the framework of the INTERPOLAR project, encompassing a total of 13 locations within the domain of university medicine in Germany. A preliminary evaluation has revealed that MRP was detected in the expected range at all 13 locations.

Discussion

The CDS Tool Chain, including the submodule 'MRP-Calculation', is both available to the public and open source. The publication of the lists of contraindications is being planned at the moment. The MRP Calculation contains measures to enable calculations to be carried out even when data is missing. The quality of the FHIR-based initial data is, nevertheless, a significant factor.

Due to the modular structure, almost any additional analysis can be added.

[29-P1-28] Real-World Interprofessional Instant Messaging in Inpatient Care: A Mixed-Methods Study in Psychosomatic Medicine

Presenter: Sabine Sayegh-Jodehl

Authors: Sabine Sayegh-Jodehl [1]

Affiliations: [1] Charité Universitätsmedizin Berlin, Berlin

Introduction

Secure messaging is becoming an important component of health information systems for inpatient coordination, yet evidence on its routine clinical use based on authentic communication data remains limited, especially in German-speaking settings. We examined how interprofessional instant messaging supports clinical coordination in a university hospital psychosomatic inpatient service and derived implications for the evaluation, design, and governance of secure messaging infrastructures.

Methods

We conducted a mixed-methods study using anonymized routine data from an on-premise Mattermost-based communication platform. The corpus comprised 410 conversations (302 direct message threads, 108 group conversations) and 7,540 messages collected between August 2021 and January 2023. Quantitative analyses included conversation structure, message length, speaker counts, temporal usage patterns, emoji frequencies, sentiment classification, and topic modeling with BERTopic. Qualitative content analysis was applied to a purposive subsample of group conversations, and direct-message questions were categorized using a rule-based approach.

Results

Messaging was predominantly instrumental and coordination-oriented, with clear modality-specific functions. Direct messages were typically short and used for dyadic micro-coordination, clarification, and negotiation of responsibilities. Group conversations were longer and concentrated on team-level handovers, alignments, requests, and announcements. Communication activity clustered during weekday core working hours direct messages peaked in the late morning, whereas group communication peaked in the early afternoon. Sentiment was predominantly neutral in both modalities. Topic modeling identified clinically related and organizational themes: direct messages covered a broader thematic spectrum, whereas group communication focused more strongly on handover-relevant content and shared situational awareness. In direct messages, most identified question types reflected coordination needs rather than knowledge seeking.

Discussion

Real-world messaging data show that different communication modalities support distinct coordination functions in inpatient care. The findings can inform health information systems research and practice by highlighting how secure messaging tools should support clear channel conventions, task ownership, escalation pathways, and handover-oriented communication design. Beyond psychosomatic medicine, the study provides transferable evidence for the evaluation and governance of digital communication infrastructures in coordination-intensive inpatient settings.

[29-P1-29] Process Support for Violence Protection Concepts: Findings from Elderly and Special Needs Care

Presenter: Johannes Schobel

Authors: Miriam Kappe [1]; Tobias Hermann [1]; Johannes Schobel [1]

Affiliations: [1] DigiHealth Institute, Neu-Ulm University of Applied Sciences, Neu-Ulm

Introduction

An inherent aspect of working in elderly and special needs care is to protect clients from violence, abuse and mistreatment. However, there is no standard process for documenting, handling, or providing aftercare for such incidents. Such processes have not yet been considered in detail in violence protection concepts.

Methods

To address this gap, a qualitative study involving 18 interviews with violence protection officers and managers from 16 institutions was conducted. The objective is to provide digital guidance for violence protection concepts. Relevant sub-processes that can be digitally supported were identified and discussed in detail with experts. All interviews were analyzed using Kuckartz's content analysis method.

Results

Domain experts described the current process as follows: An incident is reported via telephone, email, or paper-based forms. There is still no uniform documentation of violent incidents across the institutions involved. Next, the case is processed. Clarifying the facts requires a lot of communication, which should be documented. Clear guidelines are often lacking, multiple employees are involved, and manual tasks that are not supported by digital applications need to be performed. Involved participants need access to the documentation at any time. Also, there is no clear oversight of the type, extent, and risk of violent incidents. In serious incidents, authorities and other stakeholders must be involved. This is often not supported through digital processes. Subsequent process steps include transferring the data for threat/risk assessment, rehabilitating wrongly suspected individuals, providing psychological support to those affected, as well as self-reflection and prevention.

Discussion

The results clearly showed that a process-oriented approach to violence protection and to the documentation and handling of incidents is demanded by experts. Modeling, developing and executing processes can contribute to better task coordination and thus to significantly faster response times and comprehensive handling of incidents. Also, critical information can be shared with involved individuals, dramatically reducing communication overhead. A comprehensive understanding of the problem, the implementation of measures, and the analysis of their effectiveness would be possible. The more precise assessment of threats and risks contributes specifically to greater protection for clients and employees.

[29-P1-30] Refinement of Knowledge Artifacts for Molecular Tumorboards: FHIR Profiling, Visualization and AI-Assisted Information Extraction

Presenter: Kevin Kaufmes

Authors: Kevin Kaufmes [1]; Sina Thürwächter [1]; Stephanie Berger [2]; Christian Fegeler [1]; Stefan Sigle [1]

Affiliations: [1] MOLIT Institute gGmbH, Heilbronn; [2] SLK Kliniken Heilbronn, Heilbronn

Introduction

Systematic reuse of relevant information remains challenging for molecular tumor boards (MTBs) [1]. Prior work showed the feasibility of extracting Knowledge Artifacts (KA) from MTB case data and related free-text comments [2]. Limitations included missing contextual nuances from naive regex-based extraction, absence of a user interface, and KAs not being leveraged for local decision support.

Methods

This work extends our existing FHIR Implementation Guide, adding an interactive visualization layer and implementing a Large Language Model (LLM) based extraction from MTB cases. To overcome

regex limitations, a local LLM (Gemma3:27B) was integrated as an alternative method, targeting literature references in free-text comments. A KA is only created when at least one literature or clinical trial recommendation is attached to the case. A frontend web application (Vue.js) enables clinicians to search and explore KAs by gene, condition, and protein change.

Results

Application of the pipeline to the full dataset yielded 820 KAs for 1791 completed MTB cases. The extended IG successfully represents therapy recommendations, associated substances and user-specific bookmarks. The frontend application provides interactive navigation and filtering of the KA repository.

Discussion

LLM-based extraction demonstrates promising handling of contextual nuances in free-text comments compared to the regex approach. Initial evaluation revealed hallucinated literature references making post-extraction validation a necessity. Through continuous KA generation, a growing, structured knowledge base can be built. By design, KAs ensure patient privacy through abstraction from the MTB case. Future work will focus on systematic post-extraction validation, direct comparison of both extraction approaches, and leveraging the KA repository as a knowledge source for integrated clinical decision support within the MTB workflow.

1. Bartels M, Chibaudel B, Dienstmann R, Lehtiö J, Piccolo A, Michielin O, u. a. Evolving and Improving the Sustainability of Molecular Tumor Boards: The Value and Challenges. Cancers. 20. August 202416(16):2888. doi:10.3390/cancers16162888 PubMed PMID: 39199658 PubMed Central PMCID: PMC11353190.

2. Kaufmes K, Klaumünzer M, Vladimirova D, Fegeler C, Sigle S. Design and Implementation of Knowledge Artifacts to Support Molecular Tumorboard Preparation Processes. Gesundh - Gem Koop Dtsch Ges Für Med Inform Biom Epidemiol GMDS. 2024. doi:10.3205/24GMDS053

P2 - Prediction and Personalized Medicine

Poster area: P2

[29-P2-01] Benchmarking Class Imbalance Handling Strategies in Machine Learning for Sepsis Mortality Prediction: Discrimination Performance and Clinical Utility

Presenter: Jean-Pierre Gnimatin

Authors: Jean-Pierre Gnimatin [1, 2]; Siegbert Rieg [3]; Lennart Purucker [4, 5]; Philipp Mathé [3]; Martin Wolkewitz [1, 2]

Affiliations: [1] Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg im Breisgau; [2] Freiburg Center for Data Analysis, Modeling and AI, University of Freiburg, Freiburg im Breisgau; [3] Division of Infectious Diseases, Department of Medicine II, Medical Faculty and Medical Center - University of Freiburg, Freiburg im Breisgau; [4] Machine Learning Lab, University of Freiburg, Freiburg im Breisgau; [5] Prior Labs, Freiburg im Breisgau

Introduction

Machine learning (ML) models are increasingly used for clinical prediction tasks, including sepsis mortality prediction. However, real-world clinical datasets often exhibit class imbalance, which can bias model training and limit clinical usefulness. Although various imbalance-handling strategies have been proposed, their comparative impact across multiple ML algorithms in sepsis mortality prediction remains insufficiently characterized. This study aimed to benchmark different data balancing strategies and assess their impact on both discriminatory performance and clinical utility in ML models for sepsis mortality prediction.

Methods

Data were obtained from the Invasive Staphylococcus aureus Infections Cohort (INSTINCT), a prospective multicenter cohort conducted in two German university hospitals between 2006 and 2012. The dataset included 1,354 adult patients with bloodstream infections, with Day-30 mortality as the target outcome. Ten supervised ML algorithms were evaluated under five imbalance-handling strategies: none (reference), random undersampling, SMOTE, class weighting, and EasyEnsemble. Models were trained using repeated stratified nested 5-fold cross-validation with grid-search hyperparameter tuning. Performance was assessed using AUC-ROC, sensitivity, specificity, accuracy, precision, and F1-score. Strategy "champion" models were identified and evaluated using Decision Curve Analysis (DCA) to estimate clinical net benefit across threshold probabilities from 0.05 to 0.50.

Results

SMOTE-based approaches achieved the highest discrimination for tree-based ensemble models, with peak AUCs observed for XGBoost (0.848 ± 0.028), LightGBM (0.847 ± 0.028), Random Forest (0.845 ± 0.030), and CatBoost (0.844 ± 0.031). However, these models showed only moderate sensitivity (0.44-0.49). Decision curve analysis showed that class-weighted Logistic Regression (AUC = 0.838 ± 0.02, sensitivity ≈ 0.72, specificity ≈ 0.79) achieved the most consistent clinical utility, maintaining the highest net benefit beyond a threshold probability of approximately 0.155.

Discussion

These findings reinforce that AUC alone is insufficient to judge model usefulness in clinical settings, where sensitivity and overall clinical utility are critical. Strategies such as SMOTE may also introduce artificial patient profiles that do not fully reflect physiological reality, potentially limiting generalizability in heterogeneous clinical populations. In contrast, class-weighted Logistic Regression demonstrated more balanced performance and the highest clinical net benefit across a wide range of decision thresholds, highlighting the potential advantages of simpler, well-calibrated models when working with structured clinical datasets. This highlights the importance of systematically

evaluating imbalance-handling strategies and incorporating decision-analytic metrics when developing ML-based clinical decision support tools.

[29-P2-02] Development of Prediction Models for Discrete Ordinal Outcomes in Musculoskeletal Surgery: A Simulation Study and Applied Methodological Comparison

Presenter: Thomas Stojanov

Authors: Thomas Stojanov [1, 2]; Maarten van Smeden [3]; Maaike Gademan [4]; Michael Oyewale [5]; Ruud Selles [6]; Ewout Steyerberg [7]

Affiliations: [1] University Hospital Basel, Basel; [2] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht; [3] University Medical Center Utrecht, Utrecht; [4] Department of Orthopaedics, Leiden Universitair Medisch Centrum, Leiden; [5] Research and Development, Schulthess Klinik, Zurich; [6] Department of Plastic, Reconstructive and Hand Surgery, Erasmus MC, Rotterdam; [7] University Medical Center Utrecht, Utrecht University, Utrecht

Introduction

Outcomes, such as pain are usually described on a 0-10 scale. Ubiquitous in musculoskeletal surgery literature, these outcomes often show ceiling effects postoperatively. In practice, they are modelled using a wide range of statistical and machine learning approaches. However, it remains unclear how these modelling choices affect predictive performance under realistic distributional conditions. Empirical guidance remains limited. We therefore aimed to compare the performance of different modelling strategies for discrete ordinal outcomes under varying outcome distributions, sample sizes, and prognostic relationship strength, using both simulated and empirical data.

Methods

We established distinct simulation scenarios, where outcome distributions were simulated using normal distributions with varying (1) means (μ=5, 10) (to induce ceiling effects) and (2) dispersion levels (σ=1.5, 3) a rounding process was applied to discretize the outcome on a 0-10 scale (3) two sample sizes (N=100, 5000) and (4) two levels of prognostic relationship strength for a continuous predictor X (Strong, where σnoise=0.8σ vs. Moderate, where σnoise=1.75σ). X was thought as a single continuous variable representing the comprehensive predictor information. Ordinary least square (OLS), Tobit, Ordinal, Multinomial, Random Forest (RF) and XGBoost (XGB) models were compared in their ability to generate point predictions across 100 repetitions using mean root mean squared error (RMSE) and calibration curves. The estimand was the predicted Y after application of a rounding processing step and truncation at 0 and 10 where necessary. Findings were assessed using four Swiss and Dutch databases of patients with musculoskeletal disorders undergoing various surgical interventions, such as hand surgery (Hand Wrist Study Group data), arthroscopic rotator cuff repair (ARCR_Pred data), total hip or knee arthroplasty (SIRIS data).

Results

In normal outcome distributions (μ=5), mean RMSE was similar across regression strategies (1.25 for strong predictors 2.5 for moderate). Machine learning models showed lower RMSE (RF = 0.75), particularly XGB at small sample sizes (N=100, RMSE = 0.55 vs. 1.21 at N=5000). With ceiling effects (μ=10) and strong predictors, OLS overestimated the lower distribution. In that context, Multinomial and Ordinal models outperformed OLS (mean RMSE 1.52 vs. 1.60). Applied data showed similar patterns, with clinical area and model specification mainly driving differences in predicted outcome distributions.

Discussion

Standard OLS models fail to adequately predict outcomes in presence of ceiling effects in the development sample. Given the widespread interest of prediction for patient-reported outcome measures (PROMs) with skewed distributions, model choice should closely align with the underlying outcome structure.

[29-P2-03] Predicting utility scores - are standard linear regression models adequate?

Presenter: Dan Jackson

Authors: Dan Jackson [1]; Fanni Zhang [2]; Di Ran [2]; Michael Sweeting [3]; Robert Hettle [3]

Affiliations: [1] Statistical Innovation, AstraZeneca, Cambridge; [2] AstraZeneca, Gaithersburg; [3] AstraZeneca, Cambridge

Introduction

Health utility measurement is critical for health economic evaluations involving cost utility analysis. Utility mapping refers to the development and use of mathematical algorithms to predict patient health-state utility values using other measurements of health. Statistical regression models are commonly used to develop this type of algorithm. For example, regression-based methods to predict the Euro Quality of Life 5 Dimensions 3 Level Version (EQ-5D-3L) utility values have received considerable attention because their use is supported by the National Institute for Health and Care Excellence (NICE). However, these, and other, utility scores have properties that make linear regression unsuited to utility mapping. For example, values are bounded at one (full health), and distributions are often skewed and/or multimodal. More sophisticated regression models may, therefore, be considered likely to improve prediction performance.

Methods

We perform a case study to investigate the use of a variety of alternative regression models that progressively relax these assumptions. Our case study illustrates the use of a new three-step statistical framework for assessing the impact of accommodating ceiling effects, skewness, and multimodality.

Results

We find that covariate selection is a more important statistical consideration than the precise form of the regression model. However, this conclusion is only made possible by considering a range of alternative modelling approaches, including mixture models designed to accommodate ceiling effects, skewness, and multimodality.

Discussion

We recommend carefully and judiciously selecting the covariates used for prediction and, if linear regression is to be considered, comparing its performance to more flexible statistical models. We also discuss the use of regression methods for categorical outcomes within our proposed framework.

[29-P2-04] Interpretable Kolmogorov-Arnold Networks via Penalized Splines for Clinical Prediction Modeling

Presenter: Ester Rosa

Authors: Ester Rosa [1, 2, 3]; Stefania Lando [1, 2, 4, 5]; Gloria Brigiari [6, 2, 4]; Dario Gregori [6, 2]

Affiliations: [1] Unit of Biostatistics, Epidemiology and Public Health, Department of Cardiac, Thoracic, and Vascular Sciences and Public Health, University of Padova, Padova, Italy, Padova; [2] BIOSTAT-X Biostatistics & AI for Biomedical Discovery, Institute of Pediatric Research (IRP) 'Città della Speranza', Padova, Italy, Padova; [3] PhD Program in Translation Specialistic Medicine "G.B. Morgagni", Curriculum "Biostatistics and Clinical Epidemiology", University of Padua, Padova; [4] PhD Program in Translation Specialistic Medicine "G.B. Morgagni", Curriculum "Biostatistics and Clinical Epidemiology", University of Padua, Italy, Padova; [5] Department of Surgery, Oncology and Gastroenterology (DiSCOG), University of Padova, Padova, Italy, Padova; [6] Unit of Biostatistics, Epidemiology and Public Health, Department of Cardiac, Thoracic, and Vascular Sciences and Public Health, University of Padova, Padova, Italy, Turbigo

Introduction

Kolmogorov-Arnold Networks (KANs) have recently emerged as an alternative neural network architecture in which scalar edge weights are replaced by learnable univariate functions. This formulation offers a promising compromise between predictive flexibility and interpretability, a central challenge in clinical biostatistics and medical data science, particularly in small- to moderate-sized datasets. We propose a penalized spline formulation of KANs that embeds them within a generalized regression framework and provides a principled statistical characterization of model complexity.

Methods

KANs were formulated using B-spline expansions for the univariate edge functions, with difference penalties imposed to control smoothness. Model estimation was carried out via penalized log-likelihood within an iteratively reweighted least squares framework. This yields a block-structured

regularization matrix and an approximate expression for effective degrees of freedom (DoF), defined as the trace of a generalized influence operator. The approach was evaluated through a simulation study with binary outcomes, 11 standardized predictors, and four data-generating regimes of increasing structural complexity (linear, moderate, high, extreme nonlinearity). Predictive performance was assessed over 100 replications using test-set AUC and compared with logistic regression and a standard feed-forward neural network.

Results

In the linear setting, logistic regression performed best, reflecting correct model specification. As nonlinearity increased, its performance deteriorated, while spline-based KANs achieved progressively higher predictive accuracy. In highly and extremely nonlinear scenarios, KANs outperformed both logistic regression and conventional neural networks. The effective DoF provided an interpretable, continuous measure of model flexibility, linking spline basis dimension, network width, and penalty strength. Importantly, the learned univariate functions remained directly inspectable, revealing clinically meaningful patterns such as thresholds, saturation effects, and non-monotonic relationships.

Discussion

Spline-based KANs can be interpreted as a structured extension of additive regression models that bridge generalized linear models and neural networks. Their penalized likelihood formulation enables explicit regularization, interpretability, and principled complexity control. These properties make KANs a promising methodological contribution for transparent and flexible prediction modeling, supporting the ongoing integration of AI methods into clinical biostatistics. In particular, they enable a shift from parameter-based modeling to function-based representations, aligning with emerging paradigms in interpretable machine learning and facilitating adoption in clinical decision-making contexts.

[29-P2-05] Advancing mixed-effects random forests for predictions based on multi-cohort data

Presenter: jiumeng zhang

Authors: jiumeng zhang [1]; Jordan Behrendt [2]; Tanja Vrijkotte [3]; Siyu Zhou [4]; Iris Pigeot [5]; Claudia Börnhorst [6]

Affiliations: [1] Leibniz-Institute for Prevention Research and Epidemiology - BIPS, Bremen; [2] University of Bremen, CSL, Bremen; [3] Amsterdam UMC, Amsterdam Public Health, Amsterdam; [4] Amsterdam UMC, Amsterdam; [5] Leibniz Institute for Prevention Research and Epidemiology - BIPS, Bremen; [6] Leibniz Institute for Prevention Research and Epidemiology - BIPS, Bremen

Introduction

Mixed-effects models (MEMs) are widely used in epidemiology to analyze data not being independent and identically distributed (i.i.d.) like longitudinal data. However, MEMs rely on parametric assumptions and require predefined interactions among predictors. In contrast, machine learning (ML) methods such as random forests (RF) assume i.i.d. data but are more flexible in capturing nonlinear

relationships and interactions. Mixed-effects ML (ME-ML) methods, which combine MEMs with ML, have shown improved prediction performance by leveraging the strengths while mitigating limitations of both approaches. Predicting health outcomes over long time spans is challenging due to limited age range covered within single cohorts. Pooling cohorts could extend the observation periods and may improve the predictions of individual-level health trajectories by sharing information across time. However, previous ME-ML methods remain underexplored in multi-cohort settings. We propose multi-cohort mixed-effects random forests (MERFmulti-cohort), a new approach combining MEMs and RF to improve the predictions of individual-level health trajectories beyond the actual individual measurement period based on multi-cohort data.

Methods

Mixed-effects random forests (MERF) iteratively estimate fixed effects and random effects, where fixed effects are captured by the RF while the dependencies in the data are taken into account through suitable correlation structures in the MEM. However, the covariance structures of random effects and residuals are oversimplified in previous MERF implementations and may not be suitable for multi-cohort data with complex hierarchical structures. Our MERFmulti-cohort approach builds on the out-of-the-box implementation of mixed-effects machine learning framework (mixedML) [1] and extends MERF to accommodate complex covariance structures in multi-cohort data.

Results

We illustrate and evaluate our approach by predicting individual-level body mass index (BMI) and blood pressure (BP) trajectories using harmonized multi-cohort data. We compare our MERFmulti-cohort to a baseline model using previous outcome measurements as predictions, linear regression, linear mixed models, random forest, and previous MERF methods [2]. The prediction accuracy is evaluated in scenarios predicting health trajectories either based on single or multi-cohort data.

Discussion

We provide guidance on prediction scenarios where MERFmulti-cohort outperforms other methods. We further discuss the benefits of using multi-cohort over single cohort data in enhancing the accuracy of individual-level predictions.

[1] https://openreview.net/pdf?id=MKZyHtmfwH

[2] doi: 10.1177/0962280220946080.

[29-P2-06] Multi-Criteria Evaluation of Machine Learning Classifiers for Biomarker Discovery in High-Dimensional DNA Methylation Data

Presenter: Kuldeep Kumar Sharma

Authors: Kuldeep Kumar Sharma [1]; Binukumar Bhaskarapillai [1]; Binu Vs [1]; Kuppan Gokulakrishnan [1]; Chinnasamy Thirumoorthy [1]; Thennarasu Kandavel [1]; Ponnusamy Saravanan [2]

Affiliations: [1] National Institute of Mental Health and Neurosciences, Bengaluru; [2] Diabetes, Endocrinology & Metabolism University of Warwick, Warwick

Introduction

DNA methylation (DNAm) is a key epigenetic modification, and datasets capturing DNAm are typically high-dimensional. Among the large number of predictors, only a small subset is truly informative, making accurate identification for disease outcome prediction highly challenging. Although several studies have evaluated classifiers such as Random Forest (RF), Support Vector Machines (SVM), Least Absolute Shrinkage and Selection Operator (LASSO), k-Nearest Neighbors (kNN), and Naïve Bayes (NB) in high-dimensional settings, a rigorous multi-criteria framework for assessing their performance, particularly in DNAm data, is lacking. This study addresses this gap by proposing a structured multi-criteria evaluation framework tailored for DNAm datasets.

Methods

Epigenome-wide DNAm data were obtained from the STRiDE (STratification of Risk of Diabetes in Early pregnancy) prospective cohort, comprising 258 pregnant women (Normal: n = 146 GDM: n = 112). Machine-learning classifiers RF, SVM, LASSO, Logistic Regression (LR), kNN, and NB were compared in terms of discrimination, classification accuracy, and calibration. Multi-criteria aggregation was performed using the TOPSIS method, and model stability was assessed through 1,000 bootstrap resamples.

Results

RF outperformed all classifiers across most evaluation metrics, achieving the highest AUC (0.964), AUPRC (0.959), and accuracy (0.895), along with optimal calibration. SVM (AUC = 0.948) and LASSO (AUC = 0.946) performed competitively, while LR, kNN, and NB showed lower performance. TOPSIS results confirmed RF's superiority with the highest overall score (0.896). Bootstrap validation further demonstrated RF's stability with minimal variability across metrics. These findings establish RF as a robust choice for biomarker panel discovery in high-dimensional methylation data. These findings established RF as a robust and reliable choice for biomarker panel discovery in high-dimensional methylation data.

Discussion

This study evaluated commonly used classifiers using a structured multi-criteria framework incorporating discrimination, classification accuracy, and calibration. Random Forest demonstrated the best overall performance across all major metrics, supported by its ensemble-based robustness and stability under bootstrap validation. Support Vector Machines and LASSO showed competitive performance, whereas kNN and Naïve Bayes performed relatively poorly in the high-dimensional setting. Overall, Random Forest emerges as a reliable and robust approach for high-dimensional DNA methylation-based biomarker development.

[29-P2-07] Strategies for preprocessing MALDI-TOF mass spectrometry data to predict antibiotic resistance using machine learning

Presenter: Raphael Koch

Authors: Raphael Koch [1]; Niklas Wiesmann [2]; Frieder Schaumburg [2]; Dominic Enders [1]

Affiliations: [1] Institute of Biostatistics and Clinical Research, University of Münster, Münster; [2] Institute of Medical Microbiology, University Hospital Münster, Münster

Introduction

MALDI-TOF mass spectrometry is widely used for microbial identification and has further demonstrated its potential for rapid antimicrobial resistance (AMR) prediction. Predictive performance and transferability of ML models, however, strongly depend on the preprocessing of MALDI-TOF data. In particular, binning and calibration strategies can influence robustness with respect to different combinations of bacterial species and antimicrobials.

Methods

We systematically evaluated different ML pipelines for MALDI-TOF mass spectra to predict AMR across multiple bacterial species. For instance, our database includes approximately 10,000 E. coli spectra from clinical routine at the University Hospital Münster. The pipelines include various preprocessing strategies and different classifiers, e.g., eXtreme Gradient Boosting, Random Forest, and Elastic Nets, as well as hyperparameter tuning. We compared three binning approaches: linear, dynamic, and a novel approach called exponential binning. Spectra were aligned with three different calibration options (none, internal, external), each using different sets of species-specific reference peaks to determine linear warping functions.

All preprocessing parameters were learned on training data only and evaluated using nested 10x5 cross-validation. External validation was performed on the DRIAMS-A dataset, using MALDI-TOF data from a different time and location. Model performance was evaluated using the area under the receiver operating characteristic curve (AUC) for various antimicrobial and species combinations.

Results

Calibration of mass spectra improved predictive performance for all selected species-antimicrobial combinations considered, likewise in the external validation setting. However, its effect is dependent on the bin size across the 2,000-20,000 m/z range and is more pronounced for small bin sizes, i.e., for high numbers of features. The use of dynamic and exponential binning led to higher AUCs than standard linear binning, particularly when reducing the number of spectral features. Notably, this reduction did not result in relevant performance loss.

Discussion

Preprocessing of MALDI-TOF data is an important step in developing ML models for AMR prediction. Calibration can enhance predictive performance and external generalization. Dynamic and exponential binning improve predictive performance while enabling dimensionality reduction. Our findings improve the development of robust ML models for clinical AMR prediction using MALDI-TOF data by identifying optimal preprocessing strategies and quantifying their impact on model performance.

[29-P2-08] Modeling genetic tumor progression in melanoma via mixture models of oncogenetic trees

Presenter: Birte Hellwig

Authors: Birte Hellwig [1]; Alpaslan Tasdogan [2]; Jan-Malte Placke [2]; Georg C. Lodde [2]; Antje Sucker [2]; Eva Hadaschik [2]; Dirk Schadendorf [2]; Klaus Griewank [2]; Alexander Rösch [2]; Selma Ugurel [3]; Jörg Rahnenführer [4]

Affiliations: [1] Institute for Medical Informatics, Biometry and Epidemiology, University of Duisburg-Essen, Essen; [2] Department of Dermatology, University Hospital Essen, Essen; [3] Bielefeld University, Medical School and University Medical Center OWL, Department of Dermatology, Venereology and Allergology, Bielefeld; [4] Department of Statistics, TU Dortmund University, Dortmund

Introduction

Cutaneous melanoma is a highly aggressive skin cancer with rising incidence. Its progression from melanocytes to metastatic disease involves complex genetic alterations. Early somatic mutations in key signaling pathway genes, such as BRAF, NRAS, and NF1, are known to be activating and promote uncontrolled cell proliferation [1]. Progression is marked by additional mutations affecting telomerase activation, cell cycle regulation, and apoptotic resistance, contributing to further tumor development and aggressiveness.

Methods

In the prospectively collected multicenter ADOREG-TRIM real-world study cohort of 552 patients with non-resectable metastatic cutaneous melanoma, the patients' tumor tissue samples were analyzed for COSMIC-annotated driver gene mutations by 29-gene panel sequencing (miSeq, Illumina). Based on the 19 genes mutated in at least 1% of the tumors, we estimated the path of tumorigenesis using an oncogenetic tree mixture model [2] with two non-trivial components describing typical genetic tumor progression paths and an additional noise component capturing mutations not following a structured evolutionary path.

Results

Overall, 121 distinct mutation patterns were observed. The patterns of 384 tumors (70%) were consistent with at least one of two progression paths identified by the statistical model. The two non-trivial components essentially differentiated between tumors with BRAF and NRAS mutations. For most mutation patterns, a clear assignment to one of the three tree components could be made. While 99 (82%) of the patterns were assigned solely to the noise component, the majority of these patterns were extremely rare, with 94 patterns observed in four or fewer tumors.

Discussion

Tumor progression in melanoma can be described by an oncogenetic tree mixture model. The estimated model is consistent with previous findings that BRAF and NRAS are initial mutation events in tumorigenesis. However, in this model, 30% of tumors did not follow the two typical progression patterns they only fit into the noise component. This suggests that while most melanomas follow

distinct, recurrent progression patterns, a considerable proportion does not, indicating substantial heterogeneity in tumor evolution.

[1] Shain AH, Bastian BC. From melanocytes to melanomas. Nat Rev Cancer. 2016 Jun16(6):345-58.

[2] Beerenwinkel N, Rahnenführer J, Däumer M, et al. Learning multiple evolutionary pathways from cross-sectional data. J Comput Biol. 200512(6):584-98.

[29-P2-09] Prognostic Value of Metabolites for Kidney Failure in Chronic Kidney Disease Patients

Presenter: Elena Butz

Authors: Elena Butz [1]; Inga Steinbrenner [1]; Ulla T Schultheiss [1]; Charlotte Behning [2]; Harald Binder [3]; Helena Hansmann [4]; Wolfram Gronwald [5]; Peter J. Oefner [5]; Elke Schaeffner [6]; Kai-Uwe Eckardt [7, 8]; Anna Köttgen [9, 10]; Peggy Sekula [9]

Affiliations: [1] Institute of Epidemiology and Prevention, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg; [2] Institute for Medical Biometry, Informatics and Epidemiology, University Hospital Bonn, Bonn; [3] Institute of Medical Biometry and Statistics (IMBI), Faculty of Medicine and Medical Centre - University of Freiburg, Freiburg; [4] Department of Nephrology, Universitätsklinikum Regensburg, Regensburg; [5] Institute of Functional Genomics, University of Regensburg, Regensburg; [6] Institute of Public Health, Charité-Universitätsmedizin Berlin, Berlin; [7] Department of Nephrology and Medical Intensive Care, Charité-Universitätsmedizin Berlin, Berlin; [8] Department of Nephrology and Hypertension, Friedrich-Alexander Universität Erlangen-Nürnberg, Erlangen; [9] Institute of Epidemiology and Prevention, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg im Breisgau; [10] Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Baltimore

Introduction

The progression of chronic kidney disease (CKD) varies substantially between individuals, making accurate risk prediction a clinical challenge. In this study, we investigated whether metabolomics data could improve prediction of kidney failure (KF) in patients with CKD. Data came from the German Chronic Kidney Disease (GCKD) study, a prospective cohort with a median follow-up of 6.5 years.

Methods

We analyzed 4,886 participants (465 progressed to KF) with available measurements for 1,096 plasma and 1,129 urine metabolites. Using the CoxBoost algorithm, a componentwise boosting method for Cox regression, we identified metabolites that added prognostic information beyond four established clinical prognostic factors: age, sex, urinary albumin-to-creatinine ratio, and estimated glomerular filtration rate (eGFR). Bootstrap-crossvalidated time-dependent area under the receiving operator characteristic curve (AUC) and Brier score (prediction error curve) were used to evaluate predictive performance.

Results

The selection process identified 46 prognostic metabolites, several of which have been previously associated with kidney function, proteinuria, or CKD. Adding these to the clinical model modestly improved KF prediction, particularly for short-term outcomes. The AUC at 2 years increased from 0.85 to 0.89 (ΔAUC=0.046 95% confidence interval [CI]: 0.017-0.076). While discrimination performance for both models with and without metabolites was lower at more distant prediction horizons, the gain remained significant at 6 years, with an AUC increase from 0.84 to 0.87 (ΔAUC=0.022 95% CI: 0.011-0.033). Notably, no statistically significant improvements were observed for the Brier score.

Discussion

While the added metabolites improve kidney failure risk prediction, the added value is likely too small to justify the high cost and complexity of metabolite profiling in routine care. We concluded that metabolites carry prognostic information for the progression to kidney failure in CKD patients. However, their added value over established risk factors appears limited in terms of clinical utility. Nevertheless, the identified metabolites may help elucidate metabolic pathways involved in CKD progression and contribute to the biological understanding of the disease.

[29-P2-10] Objective Biomarkers Associated with Frailty: An Exploratory Analysis

Presenter: Takumi Irie

Authors: Takumi Irie [1]; Ayano Takeuchi [2]

Affiliations: [1] Chuo University, Graduate School of Science and Engineering, Chiba; [2] Chuo University, Tokyo

Introduction

Frailty, a reversible preclinical condition, requires early detection and intervention. In clinical practice, screening tools are used for early identification of frailty. However, existing tools rely on self-reported data with limited reproducibility and objectivity. Consequently, there is a growing need for objective screening frameworks that integrate multidimensional biomarkers, including the gut microbiome, to enhance the reliability of frailty assessment. This study aimed to explore factors associated with frailty using machine learning, focusing on objective indicators such as biomarkers including the gut microbiome.

Methods

This study used cross-sectional data from 798 older adults aged 65 years or older residing in Kyoto Prefecture. The Frailty Index was dichotomized at 0.25 and used as the outcome variable. As explanatory variables, 239 objective indicators were used, excluding self-reported variables and those included in the Frailty Index. Feature selection was conducted using both L1-regularized (Lasso)

logistic regression and Random Forest-based importance, with selection stability evaluated via repeated 5-fold CV. Furthermore, non-linear associations were characterized using Generalized Additive Models (GAM), while the overall directionality and interpretability of associations were confirmed via multivariable logistic regression.

Results

After excluding missing values, 657 participants were included in the final analysis. The selected features were stably identified with frequencies exceeding 80% in 5-fold cross-validation. Random Forest model demonstrated superior predictive performance (AUC: 0.72) compared to L1-logistic regression (AUC: 0.62), suggesting the presence of complex interactions among objective biomarkers. LDL cholesterol, waist circumference, and the bacterial family Lactobacillaceae were associated with frailty. In addition, C-peptide, reflecting insulin secretion, and cystatin C, a marker of renal function, showed non-linear associations with frailty.

Discussion

This study exploratorily identified characteristics of objective indicators associated with frailty. No single factor emerged multiple disease-related indicators, such as chronic kidney disease and abnormalities in glucose metabolism, were consistently associated with frailty. While the model achieved moderate discriminative ability, the performance was lower than benchmarks set by self-reported tools, indicating that biological markers capture a distinct dimension of frailty that may not fully overlap with functional perceptions. These findings suggest that objective indicators alone may not fully capture frailty status, and adding simple questionnaire items may improve performance.

[29-P2-11] Smoking-stratified performance of a polygenic score for lung adenocarcinoma in an East-Asian employee cohort

Presenter: Shunsuke Minusa

Authors: Shunsuke Minusa [1]; Hideyuki Ban [1]; Takeshi Tanaka [1]; Yuya Watanabe [2]; Tohru Nakagawa [2]

Affiliations: [1] Research & Development Group, Hitachi, Ltd., Kokubunji; [2] Hitachi Health Care Center, Hitachi, Ltd., Hitachi.

Introduction

Current lung-cancer screening protocols, which primarily rely on age and smoking history, potentially overlook high-risk non-smokers. Although polygenic scores (PGS) provide valuable insights for precision prevention, their effectiveness may be obscured by environmental influences. This study preliminarily evaluated the difference of a 28-single nucleotide polymorphism (SNP) PGS for lung adenocarcinoma in a high-fidelity East-Asian employee cohort, specifically investigating how smoking status modulates the genetic signal.

Methods

We conducted a retrospective case-control study using an East-Asian employee cohort (n=48 12 cases, 36 controls) as a pilot study based on an extreme phenotype strategy. Cases were lung-cancer survivors (69.2±4.4 years), and controls were highly healthy individuals without lung cancer history and abnormal findings in annual checkup results chronologically (68.6±4.9 years). SNP genotyping was performed using Illumina Asian Screening Array, with imputation by Beagle 5.5 (1000 Genomes Phase 3 v5 as reference). We calculated a 28-SNP PGS based on published log-odds weights (QC: INFO≥0.5, MAF≥0.01, missing≤2%, HWE $p \geq 1 \times 10^{-6}$) [1]. PGS differences were assessed using AUC and Mann-Whitney U tests for both the total population and smoking-stratified subgroups. All procedures were in accordance with the Declaration of Helsinki, and the study was approved by the ethics committee of Hitachi group headquarters (#334-2).

Results

In the non-stratified analysis, PGS differences were negligible (AUC=0.677 p=0.068 when applying non-smokers PGS) However, stratification revealed significant signal amplification among non-smokers (cases n=5, controls n=16) cases exhibited significantly higher PGS (p=0.001), achieving an AUC of 0.950 (95% CI 0.843-1.000). Conversely, no significant discrimination was observed in the smoking group (p=0.116).

Discussion

The findings indicate that genetic susceptibility is more detectable in the absence of tobacco exposure, supporting a critical gene-environment interaction. This study demonstrates that stratified validation, even with limited samples of extreme-phenotype datasets, can reveal high-performing genetic signals that might otherwise be diluted. These results underscore the importance of integrating lifestyle context into genomic risk models and support a shift toward risk-based low-dose computed tomography screening for genetically high-risk non-smokers. Limitations include small sample size, male-only data, and AUC calculation without cross-validation further validation in larger, more diverse populations is necessary.

[1] Shi J, et al. Nat Commun. 202314:3043.

[29-P2-12] Examining polygenic risk scores for broad depression that account for heterogeneity due to sex

Presenter: Sharon Lutz

Authors: Sharon Lutz [1]; Christoph Lange [2]

Affiliations: [1] Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston; [2] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Harvard University, Boston

Introduction

Polygenic risk scores (PRSs) estimate an individual's genetic liability to a trait or disease. In recent years, PRSs have generated substantial interest in precision medicine due to their potential to identify people who are at higher risk for certain conditions. However, PRSs may have reduced predictive performance if genetic diversity and environmental/sex effects are not considered.

Methods

Through simulation studies, we examined the predictive performance of PRSs when there is heterogeneity due to sex. We considered overall, sex stratified, and sex by SNP interactions. Since depression is a polygenic disorder (due to the combined effects of many genetic variants) and depression rates differ by sex, we used the PRS-CSx approach to examine sex specific effects. The PRS-CSx approach combines GWAS summary statistics from different ancestral groups and uses Bayesian regression with continuous shrinkage priors to adjust SNP effect sizes based on GWAS effect estimates and LD structure. We applied the PRS-CSx approach in the UK Biobank for the training data sets with testing in the Mass General Brigham Biobank accounting for heterogeneity due to sex.

Results

In the Mass General Brigham Biobank, we showed that predictive performance for the PRS can be increased by accounting for heterogeneity due to sex.

Discussion

Methods for PRS can have increased predictive power for complex traits such as depression if heterogeneity due to environmental/sex effects are considered.

[29-P2-13] A novel integrated molecular prognostic model and risk-score (IMR) combining weighted molecular genetic-risk and clinical risk-criteria in BCR-ABL1 negative pediatric B-cell acute lymphoblastic leukemia

Presenter: Vineet Kumar Kamal

Authors: Vineet Kumar Kamal [1]

Affiliations: [1] All India Institute of Medical Sciences (AIIMS), Kalyani

Introduction

To develop, validate, and assess the clinical utility of an integrated molecular prognostic model and risk score (IMR) that combines weighted molecular genetic alterations with National Cancer Institute (NCI) clinical risk criteria for improved risk stratification in pediatric BCR-ABL1-negative B-cell acute lymphoblastic leukemia (B-ALL). Methods

This study included pediatric patients with BCR-ABL1-negative B-ALL treated under the ICICLE protocol. Baseline genomic profiling was performed using multiplex ligation-dependent probe amplification (MLPA) to detect deletions in key genes, including IKZF1, PAX5, CDKN2A/B, BTG1, RB1, ETV6,

EBF1, ERG, and pseudoautosomal region (PAR) genes. Clinical variables, including age and total leukocyte count (TLC >50,000/μL), were incorporated as per NCI risk criteria. The primary outcome was time to first event (relapse or death), with event-free survival (EFS) and overall survival (OS) evaluated. A Cox proportional hazards regression model was used to estimate the probability of first event at 2 and 5 years. Variables with pA total of 284 patients were included in the development cohort, with a median age of 6 years (IQR: 3-12) and 62.7% males. High TLC (>50,000/μL) was observed in 22.9% of patients. MLPA analysis identified deletions in CDKN2A/B (31%), PAX5 (24.7%), IKZF1 (12.7%), ETV6 (12.7%), RB1 (9.9%), BTG1 (6.7%), PAR1 (4.6%), and EBF1 (2.8%). During a median follow-up of 6.3 years, 155 events occurred, including 113 deaths, with median EFS of 3.3 years. The IMR scoring system assigned 5 points each for IKZF1 and RB1 deletions, 4 points for CDKN2A/B deletion, 2 points for age >10 years, and 1 point for TLC >50,000/μL, while age 1-4 years was assigned -3 points. The model demonstrated moderate discrimination, with Harrell's C-statistic of 0.62 in the development cohort, 0.60 in internal validation, and 0.55 in external validation (n=140). Predicted EFS decreased progressively with increasing risk scores: at 2 years (68%, 53.4%, 36.1%, 19.1%) and at 5 years (49.8%, 32.3%, 15.9%, 5%) for scores of 0, 5, 10, and 15, respectively. Significant differences in EFS and OS were observed across risk groups (10 pDiscussion

The IMR model combines molecular genetic risk with NCI clinical criteria, offering a novel approach to risk stratification in pediatric BCR-ABL1-negative B-ALL. It demonstrates moderate discrimination, strong calibration, and clinical utility, making it a promising tool for personalized treatment strategies.

[29-P2-14] A Patient Similarity Network Framework for Integrating Clinical and Molecular Data in Precision Oncology

Presenter: Iryna Manuilova, Arsenij Ustjanzew

Authors: Iryna Manuilova [1]; Annemarie Weise [1]; Arsenij Ustjanzew [2]; Susann Schulze [3]; Julia Folwark [4]; Varvara Zakharchenia [4]; Cosima Strantz [5]; Benedikt Brors [6]; Jan Christoph [7]

Affiliations: [1] Junior Research Group (Bio-)Medical Data Science, Faculty of Medicine, Martin-Luther-University Halle-Wittenberg, Halle (Saale); [2] Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI), University Medical Center of the Johannes Gutenberg-University Mainz, Mainz; [3] Krukenberg Cancer Center Halle (Saale), Halle (Saale); [4] Junior Research Group (Bio-)Medical Data Science, Faculty of Medicine, Martin Luther-University Halle-Wittenberg, Halle (Saale); [5] Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen; [6] Division Applied Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg; [7] Martin-Luther-University Halle-Wittenberg, Halle

Introduction

Patient similarity networks (PSNs) are a promising approach in precision oncology because they enable data-driven stratification of patients based on clinical and molecular characteristics. They may support molecular tumor boards by identifying comparable cases and informing interpretation and treatment decisions. However, similarity modeling in oncology remains challenging because clinical data are heterogeneous, incomplete, and distributed across mixed variable types and multiple

records per patient. We therefore developed and evaluated a PSN framework for integrating heterogeneous clinical and molecular data in precision oncology.

Methods

Within the German PM4Onco (Personalized Medicine for Oncology) project, we developed a framework to construct PSNs from clinical and molecular data. Clinical and mutation-derived data were preprocessed separately, and record-level dissimilarities were computed using weighted Gower distance for mixed data types. For patients with multiple records, cross-record Gower distances were aggregated into a single patient-pair distance using the Chamfer distance. Clinical distances were calculated with equal or clinician-informed variable weights, whereas mutation distances used burden score-based weights to emphasize clinically relevant alterations. For each modality, affinity matrices were derived from the distance matrices and fused by Similarity Network Fusion (SNF) to generate the final similarity network. The framework was applied to four cancer cohorts of The Cancer Genome Atlas (TCGA) and evaluated using clustering and survival analyses.

Results

Across the TCGA cohorts, the fused similarity networks revealed distinct patient clusters. The similarity space showed clinical coherence, with comparable patients exhibiting lower pairwise distances. Comparison of community detection methods across several evaluation metrics showed substantial agreement, particularly for Louvain, Spinglass, and Fast Greedy. The identified clusters were significantly associated with overall survival and corresponded to established cancer subtypes.

Discussion

These results suggest that the proposed PSN framework can capture clinically relevant structure despite the heterogeneous, incomplete, and multi-record nature of oncology data. The combination of Gower distance, Chamfer aggregation, and SNF provides a flexible strategy for integrating clinical and molecular information within a unified similarity model. However, the effect of different weighting schemes requires further systematic evaluation. Overall, these findings support further validation of the framework in molecular tumor board workflows.

[29-P2-15] Accelerating Alzheimer's Research: a modular framework for exploring Bayesian disease progression models

Presenter: Oana Petrof

Authors: Oana Petrof [1]; Dave Lunn [2]; Aris Perperoglou [3]

Affiliations: [1] GSK, Wavre; [2] GSK, UK, London; [3] GSK, London

Introduction: Alzheimer's disease (AD) is a progressive neurodegenerative disorder characterized by considerable heterogeneity in clinical progression. Building upon the framework proposed by Lars Lau Raket et al. in Statistical Disease Progression Modelling in Alzheimer Disease, we developed a statistical approach for modelling disease progression using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. Our method provides a common timescale representation of

disease progression by aligning individual cognitive trajectories. Methods: We utilize a Bayesian hierarchical model that simultaneously operates on cognitive scale responses and disease time. Cognitive decline is captured through a nonlinear mixed-effects model. Initially, covariate-free models are fit to establish posterior distributions for each patient's disease age (DA) using different cognitive endpoints such as ADAS-COG 13, CDR sum of boxes, MMSE, all assessing cognition on different scales. Our contribution is a two-stage modularization approach which decouples the computationally intensive single model estimation of DA in stage 1 from the subsequent joint endpoint modelling and the covariate analysis in stage 2. We explore different ways of combining multiple endpoints into a joint model connecting the endpoints through DA, different covariate models for each individual endpoint as well as different covariate models for each joint endpoint model. This modularization allows for the rapid exploration of multiple cognitive endpoints and covariate models utilizing demographics and biomarkers as predictors. Results: The power of this modularization design lies in its second stage: it enables the rapid and rigorous exploration of numerous, distinct joint models and covariate models without requiring the re-running of the expensive initial fit. This process formally propagates the parameter uncertainty captured in the disease age posteriors. We achieve significant efficiency gains without compromising the integrity of the inference by using our proposed modularization of complex models. Discussion: This modularization framework provides a robust method for assessing demographics/biomarker effects in the context of cognitive decline. The second-stage model can address real-world data challenges, including covariate missingness and non-normal distributions of biomarkers. By improving personalized prognostication and facilitating more effective therapeutic stratification, our methodological refinements advance the statistical modelling of AD progression.

[29-P2-16] Developing prognostic models in rare cancers: A systematic review of sample size and methodological approaches in recent studies

Presenter: Laura Kirton

Authors: Laura Kirton [1]; Richard Riley [1]; Jak Laight [1]; Piers Gaunt [1]; Susanne A Gatz [1, 2]; Kym Snell [1]

Affiliations: [1] University of Birmingham, Birmingham; [2] Birmingham Women's and Children's NHS Foundation Trust, Birmingham

Introduction

Current sample size guidance for developing prognostic models in a specific target population aims to minimise overfitting and optimism in the R2, and target precise estimation of overall risk [1]. In the rare cancers setting, reaching this minimum sample size is often unachievable yet there remains strong interest from patients and clinicians for individualised risk predictions which raises methodological challenges.

The objective of this systematic review is to examine recent prognostic model development studies conducted in rare cancers. Specifically, the review aims to:

Investigate the level of adherence to current sample size guidance [1] Summarise whether and how sample sizes were determined Highlight methodological approaches used to address sparse data during model development and Determine whether uncertainty and instability of the predictions, and clinical utility were evaluated.

Methods

Sarcomas were used as the clinical exemplar, as all sarcoma histological subtypes are classified as rare cancers. A systematic search of MEDLINE and Embase identified studies published between January 2019 to August 2024 developing prognostic models in sarcoma populations. From the eligible publications, key data were extracted: sample size characteristics, details and justification of model development methods, and any assessment of prediction instability, uncertainty and/or clinical utility. Where studies applied statistical methods to address the limited sample size, a qualitative summary was extracted.

Results

Of the 74 eligible studies, none prospectively considered or addressed the sample size, and most did not adhere to the Riley et al. sample size guidance [1]. None implemented novel methodological strategies to address small sample sizes which are inherent to the disease setting. Internal and external validation were the only downstream assessments of model instability, with the "split-sample" method often employed. Uncertainty intervals were occasionally reported with the individualised predictions, although the methods used to calculate them were not described. Decision thresholds for clinical utility assessment were rarely considered.

Discussion

The findings provide an important overview of current prognostic model development studies in rare cancers, highlighting concerns that sample size limitations are frequently overlooked. Rare cancers lead to limited available sample sizes for prognostic modelling, and therefore this review concludes with recommendations for handling these constraints.

References:

[1] doi:10.1002/sim.7992

[29-P2-17] Unsupervised Phenotyping of ICU Patients at Risk for Post-Intensive Care Syndrome Using the eICU Collaborative Research Database

Presenter: Johanna Schwinn

Authors: Johanna Schwinn [1]; Felix Berger [2]; Seyedmostafa Sheikhalishahi [3]; Manfred Weiß [2]; Philipp Simon [2]; Ludwig Christian Hinske [3]; Mathias Kaspar [1]

Affiliations: [1] Universitätsklinikum Augsburg, Augsburg; [2] Anesthesiology and Operative Intensive Care, Faculty of Medicine, University of Augsburg, Augsburg; [3] Digital Medicine, University Hospital Augsburg, Augsburg

Introduction

Post-Intensive Care Syndrome (PICS) comprises physical, cognitive, and mental impairments persisting after ICU discharge. Because routine databases rarely contain post-discharge PICS outcomes, unsupervised methods may help identify potential high-risk subgroups from in-ICU data. We apply per-hospital hierarchical clustering to multicenter ICU data to explore whether acuity phenotypes relevant to PICS risk emerge from routinely collected variables.

Methods

Data were extracted from the eICU Collaborative Research Database, restricting analyses to hospitals with ≥1,000 admissions (n=12 sites). Per hospital, continuous features (age, ICU and hospital length of stay (LOS), comorbidity burden, vasopressor days, benzodiazepine (BZD) days and dose, delirium duration, and mean RASS) were log(1+x)-transformed where skewed, imputed with IterativeImputer, and RobustScaled. IsolationForest-based outlier removal (contamination=5%) was applied prior to clustering. Hierarchical Ward linkage was applied at k=2. Cluster stability was quantified via bootstrap Jaccard similarity (20 subsamples at 80% J<0.60 flagged as unstable).

Results

Feature availability ranged from 4 to 8 variables per site. Silhouette scores at k=2 ranged from 0.181 (H252) to 0.720 (H420), with most sites showing moderate separation (0.25-0.50) bootstrap Jaccard stability was acceptable at most sites. At most sites the two clusters were interpretable as lower- and higher-acuity, the latter showing longer ICU and hospital LOS, higher BZD dose and days, greater vasopressor burden, longer delirium, and deeper sedation. Its proportion ranged from 8% to 77% of patients across sites.

Discussion

Per-hospital Ward clustering of ICU features recovers an exposure constellation consistent with established PICS risk factors. However, without post-discharge outcomes it cannot be established that the higher-acuity cluster reflects PICS risk rather than general illness severity. At most sites the higher-acuity proportion fell below commonly reported PICS prevalence, suggesting that in-ICU exposures alone under-capture risk: routine data omit established predictors such as pre-existing cognitive impairment, frailty, and negative ICU experience. Site-dependent variation in separation and cluster proportion further reflects feature availability and per-site preprocessing, limiting cross-site comparability. Future work should evaluate alternative cluster numbers (k>2) to test whether finer, more clinically meaningful phenotypes emerge and validate against proxy outcomes. The present findings are exploratory.

[29-P2-18] Refining Post-COVID Syndrome Assessment: Distinct Symptom Domains Are Linked to Acute Disease Severity and Resilience

Presenter: Anne-Kathrin Ruß

Authors: Anne-Kathrin Ruß [1, 2, 3]; Sabrina Ballhausen-Lübcker [4]; Wolfgang Lieb [5]; Anna Schäfer [6, 7]; Lilian Krist [8]; Klaus Rabe [9, 10]; Walter Maetzler [11]; Corina Maetzler [11]; Martin Laudien [12]; Derk Frank [13]; Jan Heyckendorf [4]; Olga Miljukov [14]; Karl Georg Haeusler [15]; Nour Eddine El Mokhtari [16]; Jörg Janne Vehreschild [17]; Katharina Appel [17]; Irina Chaplinskaya-Sobol [18]; Carolin Nürnberger [19]; Lena Schmidbauer [6, 7]; Caroline Morbach [20]; Stefan Störk [21]; Jens-Peter Reese [22]; Thomas Keil [23]; Peter Heuschmann [24]; Stefan Schreiber [25]; Michael Krawczak [26]; Thomas Bahmer [25]

Affiliations: [1] Institute of Clinical Studies and Digitalization, Medical Statistics, Health and Medical University Erfurt, Erfurt; [2] Department of Human Medicine, Health and Medical University Erfurt, Erfurt; [3] Institute of Medical Statistics, Kiel University, University Hospital Schleswig-Holstein Campus Kiel, Kiel; [4] Internal Medicine Department I, University Hospital Schleswig Holstein Campus Kiel, Kiel; [5] Institute of Epidemiology, Christian-Albrechts-University Kiel, Kiel; [6] Institute of Clinical Epidemiology and Biometry, University of Würzburg, Würzburg; [7] Institute for Medical Data Sciences, University Hospital Würzburg, Würzburg; [8] Charité - Universitätsmedizin Berlin, Berlin; [9] Pneumology, LungenClinic Grosshansdorf, Berlin; [10] Airway Research Center North (ARCN), German Center for Lung Research (DZL), Berlin; [11] Department of Neurology, UKSH Kiel, Kiel; [12] ENT Department, University Hospital Schleswig Holstein Campus Kiel, Kiel; [13] Department of Cardiology, Kiel University and University Hospital Schleswig-Holstein, Kiel, Germany, Kiel; [14] Julius-Maximilians-Universität Würzburg, Würzburg; [15] Department of Neurology, University Hospital Ulm, Ulm; [16] Cardiology, Schön Klinik Rendsburg, Rendsburg; [17] Goethe-Universität Frankfurt, Institut für Digitale Medizin und Klinische Datenwissenschaften, Deutschland, Frankfurt am Main; [18] Institute for Medical Informatics, University Medical Center Göttingen, Göttingen; [19] Universität Würzburg, Würzburg; [20] Deutsches Zentrum für Herzinsuffizienz, Universität und Universitätsklinikum Würzburg, Würzburg; [21] Department Clinical Research & Epidemiology, Comprehensive Heart Failure Center Würzburg and Department of Internal Medicine I, University Hospital Würzburg, Würzburg; [22] Faculty of Health Science, Technische Hochschule Mittelhessen, University of Applied Science, Giessen, Germany, Giessen; [23] Institute of Social Medicine, Epidemiology and Health Economics, Charité - Universitätsmedizin Berlin, Berlin; [24] Universität Würzburg, Institut für klinische Epidemiologie und Biometrie, Würzburg; [25] Internal Medicine Department I, University Hospital Schleswig-Holstein (UKSH) Kiel, Kiel; [26] Christian-Albrechts-Universität zu Kiel, Kiel

Introduction

Post-COVID Syndrome (PCS) refers to the diverse long-term consequences of Sars-CoV-2 infection and subsequent COVID-19. A previously established PCS score comprising 12 symptom complexes allowed the identification of two principal predictors of PCS severity, namely acute disease severity and individual resilience [Bahmer et al. 2022]. This study aimed to validate the two predictors in a larger cohort and to examine their relationship to specific symptom complexes in more detail.

Methods

Data were obtained from 3,372 participants of the COVIDOM/NAPKON-POP study. Classification and Regression Tree (CART) analysis was used to explore the association between acute COVID-19

severity and individual resilience on the one hand, and individual symptom complexes of the PCS score on the other.

Results

The study confirmed that both, acute disease severity and individual resilience are the most significant predictors of PCS severity. In addition, distinct patterns emerged for individual symptom complexes. Thus, while neurological symptoms, sleep disturbances, and fatigue were primarily associated with resilience, acute disease severity was linked to exercise intolerance, chemosensory deficits, musculoskeletal pain, signs of infection, and fatigue. Based upon these findings, two more refined PCS scores were developed including only symptom complexes associated with the respective predictor. Consistent with the original PCS score, both new scores showed a significant inverse correlation with quality of life, as measured by the EQ-5D-5L index.

Discussion

The study validated the role of acute disease severity and individual resilience in determining PCS severity and highlighted the differential impact of these two predictors on different PCS symptom complexes. The novel PCS scores allow a more nuanced assessment of PCS severity, potentially improving clinical management by way of developing more targeted therapeutic strategies and interventions tailored to distinct PCS subdomains.

[29-P2-19] Individualized hip surveillance in cerebral palsy using latent trajectory modelling: a population-based study

Presenter: Philippe Wagner

Authors: Philippe Wagner [1]; Gunnar Hägglund [2]

Affiliations: [1] 2Centre for Clinical Research, Uppsala University, Region Västmanland, Västerås, Västerås; [2] 1Department of Clinical Sciences, Orthopedics, Lund University, Lund, Sweden, Lund

Introduction

Hip displacement is a common complication in cerebral palsy (CP) and may progress to dislocation without timely intervention. Surveillance programmes based on age and Gross Motor Function Classification System (GMFCS) level reduce dislocation risk but may not capture substantial heterogeneity in individual disease trajectories. We evaluated whether latent trajectory modelling of migration percentage (MP how much of the hip ball is outside the socket) can support individualized surveillance and reduce radiographic burden while maintaining safety.

Methods

We used longitudinal data from the Swedish CPUP registry, including 697 children in GMFCS levels III-V with 3,756 radiographic examinations before age 12. MP was defined as the maximum of both hips at each visit. Trajectories were modelled using latent class mixed models, by means of the lcmm R package (1) with cubic polynomial age terms for both fixed and random effects, including GMFCS as both a covariate and predictor of class membership. Model selection was based on the

BIC. Model adequacy was evaluated through inspection of posterior class probabilities and residual distributions. At each visit, predictions from current age to 12 years were generated by combining class-specific trajectories using empirical Bayes estimates of individual random effects. Follow-up was discontinued if posterior class probability was ≥0.80 and the upper confidence limit of the predicted trajectory remained below MP 40. Otherwise, follow-up continued according to the observed Swedish CPUP schedule. Outcomes included the number of avoided examinations and the proportion of children reaching MP ≥40 after discontinuation.

Results

Four latent trajectory classes were identified, representing distinct stable and progressive patterns. Two classes showed early increase followed by stabilization around MP 20-25, whereas two showed continued progression with higher risk of reaching MP ≥40. Under the prediction-based policy, 268/697 children (38.5%) were discontinued, reducing the number of examinations by 21.0% (788/3,756). Among discontinued children with post-discontinuation observations, 19/229 (8.3%) reached MP ≥40, corresponding to 2.7% of the full cohort. Discontinuation was concentrated in low-risk classes, while high-risk classes were largely retained.

Discussion

Latent trajectory modelling allows individualized hip surveillance in CP, identifying children with stable trajectories who may safely undergo reduced follow-up. This approach may complement existing programmes.

REFERENCES

1. Proust-Lima C, Philipps V, Liquet B. Estimation of extended mixed models using latent classes and latent processes: The R package lcmm. J Stat Softw. 201778(2):1-56.

[29-P2-20] Wavelet-based analysis of cardiovascular oscillatory dynamics in critically ill patients: the association of demographics and medications

Presenter: Jayanth Sreekanth

Authors: Jayanth Sreekanth [1, 2]; Eduardo Salgado [1, 2, 3]; Andreas Edel [1]; Alawi Luetz [1]; Lilian J. Engelhardt [1]; Henry Orlovsky [1]; Elias Grünewald [2]; Sophie K. Piper [2, 4]; Claudia Spies [1]; Felix Balzer [2]; Sebastian Boie [1, 2]

Affiliations: [1] Charité-Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Department of Anesthesiology and Intensive Care Medicine (CCM CVK), Augustenburger Platz 1, Berlin; [2] Charité-Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Institute of Medical Informatics, Charitéplatz 1, Berlin; [3] Berlin Institute of Health at Charité-Universitätsmedizin Berlin, Core Unit Digital Medicine and Interoperability, Charitéplatz 1, Berlin; [4] Charité-Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Institute of Biometry and Clinical Epidemiology, Charitéplatz 1, Berlin

Introduction

Timed and synchronized processes of the heart and vasculature are complex. Blood pressure and heart rate exhibit non-stationary oscillations that reflect their continuous, dynamic interplay within the autonomic nervous system. During critical illness, various disease-related factors and treatments influence cardiovascular oscillatory dynamics. Analyzing cardiovascular vitals from a continuous patient monitoring system requires robust analytical methods, such as wavelets, capable of capturing non-stationary signal dynamics. We use continuous wavelet transform to explore high-resolution, time-frequency resolved periodic components, and investigate how age, sex and medications administered in the intensive care unit (ICU) shape transient cardiovascular rhythms.

Methods

We study patients treated in the Charité ICUs for Anesthesiology and Intensive Care Medicine between 01 January 2017 and 31 October 2024, with a minimum stay of 5 days. Following approval by the ethics committee of Charité (EA1/284/22), continuous five-day measurements of blood pressure and heart rate were retrieved from the patient management system alongside relevant clinical data. The continuous wavelet transform was applied to blood pressure and heart rate to characterize periodic oscillations within the power spectrum. The characteristics were tracked over time using our novel oscillation tracking approach. The patients were stratified by age groups and sex to identify baseline physiological patterns. Additionally, we compared the oscillation characteristics between patients who received vasopressors, analgesics, or sedatives and those who did not receive these medications.

Results

Utilizing our wavelet-based approach, we quantified time-varying power and periodicity of blood pressure and heart rate, identifying prominent diurnal and semidiurnal oscillations over time. These oscillatory dynamics of blood pressure and heart rate with multiple periodic components revealed age- and sex-specific variations. Notably, patients with vasopressors and/or sedatives and/or analgesics showed an attenuation of intrinsic cardiovascular periodicities.

Discussion

Our findings reveal age- and sex-specific patterns, indicating the association of demographic factors on cardiovascular oscillatory dynamics. Additionally, continuous wavelet transform reveals how medications modulate the diurnal and semidiurnal oscillations of blood pressure and heart rate. By moving beyond static averages, our approach to estimating time-resolved transient oscillations supports a shift toward rhythm-aware cardiovascular monitoring.

[29-P2-21] Variational Signal Decomposition via Discrete Penalized Least Squares for Non-stationary Biomedical Time-Series

Presenter: Minsu Park

Authors: Minsu Park [1]

Affiliations: [1] Chungnam National University, Daejeon

Introduction

The analysis of biomedical signals is often challenged by non-stationary fluctuations and complex noise, which obscure essential underlying patterns. While the empirical mode decomposition (EMD) is a popular data-driven tool for such signals, its heuristic sifting process lacks a formal statistical framework. This often leads to critical limitations, including numerical instability, mode-mixing, and significant boundary artifacts, which can compromise the integrity of clinical data interpretation.

Methods

To address these limitations, we propose a variational decomposition approach by reformulating signal decomposition as a statistically rigorous discrete penalized least squares (DPLS) optimization problem. By identifying a unique global minimizer that balances data fidelity with second-order difference penalties, we characterize the estimator via the Demmler-Reinsch spectral basis. This framework establishes a deterministic mapping between the regularization parameter and the frequency response, enabling objective and reproducible control over decomposition scales.

Results

Simulation results demonstrate that the proposed method effectively separates structural trends from oscillatory components without the mode-mixing issues inherent in empirical methods. Furthermore, application to real-world biomedical data demonstrates that our method effectively removes unstable baseline fluctuations while preserving the sharp peaks and essential morphology of critical physiological events without structural distortion.

Discussion

By utilizing a sparse direct solver, the proposed framework enables efficient processing of large-scale clinical time series. Unlike iterative heuristic methods, our approach provides a unique solution and theoretical grounding. Consequently, this method serves as a robust and high-fidelity tool for biomedical signal interpretation, offering researchers a reproducible alternative for precise physiological data analysis.

[29-P2-22] Machine Learning based propofol blood concentration prediction

Presenter: Christian Schorr

Authors: Christian Schorr [1]; Thomas Jäger [2]; Lukas Müller-Wirtz [3]; Sascha Kreuer [4]

Affiliations: [1] German Research Center for Artificial Intelligence (DFKI), Saarbrücken; [2] Data Integration Centre, Medical Faculty, Saarland University, Homburg (Saar); [3] Department of Anaesthesiology, Friedrich Alexander University Erlangen, Erlangen; [4] Department of Anaesthesiology, Intensive Care and Pain Therapy, Saarland University, Homburg (Saar)

Introduction

Propofol is among the most widely used anaesthetics and is included in about 95% of surgical procedures worldwide However, safe and effective use of propofol relies on precise dosing, which in turn depends on accurate pharmacokinetic (PK) models. Machine learning offers a flexible alternative to deterministic pharmacokinetic modelling by deriving nonlinear relationships directly from data without requiring predefined compartment structures.

Methods

In a retrospective approach we evaluate the predictive performance of four machine learning methods - Random Forest, LSTM, TabPFN and XGBoost for arterial propofol plasma concentrations. As reference we also considered a three compartment NONMEM model and the Eleveld model used in most TCI pumps. Based on a pooled dataset of 4 studies containing the surgeries of 144 patients with 1,220 arterial blood samples, we evaluated each model's ability to directly predict plasma propofol concentrations from intraoperative infusion histories and patient characteristics. The dataset included both time-varying and patient-specific features. Due to the very irregularly distributed blood sampling points and different surgery lengths, extensive data pre-processing was necessary to allow for model training. Additional temporal information was embedded through engineered features such as lagged infusion rate and cumulative dose, accounting for delayed drug effects and short-term pharmacokinetic dependencies despite an irregular sampling schedule.

Results

The Random Forest model achieved predictive accuracy comparable to NONMEM (R2=0.68, MDAPE = 21.92 vs 0.69 and 22.19) but with significantly reduced bias (MDPE -1.1% vs -3.9%). The NONMEM model systematically under-predicted plasma concentrations. The other machine learning models had only slightly lower metrics than NONMEM but still significantly outperformed the standard Eleveld. A Sobol sensitivity analysis revealed that elapsed surgery time and recent infusion history were the most influential features for the Random Forest model.

Discussion

Although based on a rather small training data set, data-driven machine learning models are up to par with state-of-the-art compartiment models and show lower bias. Both outperform the clinical standard Eleveld model. Larger, more diverse datasets and more frequent measurements may further improve accuracy and generalizability. Unlike traditional PK models, ML-based models might therefore facilitate real-time patient-specific propofol dosing decisions without relying on fixed assumptions inherent in traditional compartmental models.

[29-P2-23] Predicting Umbilical Arterial Acidemia from Fetal Heart Rate Patterns: A Machine Learning Approach

Presenter: Luis Mariano Esteban

Authors: Luis Mariano Esteban [1, 2]; Javier Esteban-Escaño [3]; Antonio R. Laliena [3]; Ricardo Savirón-Cornudella [4, 5, 6]

Affiliations: [1] Escuela Universitaria Politécnica de La Almunia, Universidad de Zaragoza., Zaragoza; [2] Institute for Biocomputation and Physics of Complex Systems, University of Zaragoza, Zaragoza; [3] Escuela Universitaria Politécnica de La Almunia, Universidad de Zaragoza., La Almunia de Doña Godina; [4] Hospital Clínico San Carlos, Madrid; [5] Instituto de Investigación Sanitaria San Carlos (IdISSC), Madrid; [6] Universidad Complutense de Madrid, Madrid

Introduction

Electronic fetal monitoring (EFM) is the standard method for assessing fetal well-being during labor. This study aimed to apply machine-learning algorithms to predict neonatal acidemia using features extracted from the fetal heart rate signal within a 30-minute window.

Methods

The EFM signal was divided into deceleration and interdeceleration segments. From these variables, three parameters were derived over the last 30 minutes: total reperfusion time (sum of interdeceleration periods), total deceleration time (sum of deceleration durations), and total deceleration area. Additional features included amplitude, duration, drop, slope, area, fetal heart rate, excess deceleration, instability, and reduced deceleration. A case-control design was used, including 502 infants born at Miguel Servet University Hospital (Spain), with a 1:1 ratio of cases to controls. Neonatal acidemia was defined as an umbilical arterial pH < 7.10. Logistic regression, classification trees, random forest, and neural network models were developed using combined EFM features. Model performance was evaluated in terms of discrimination, calibration, and clinical utility.

Results

The random forest model achieved the highest AUC (0.971 95% CI: 0.948-0.993), although no statistically significant differences were observed compared with logistic regression (0.968), neural networks with one hidden layer (0.962), or two hidden layers (0.925). However, random forest significantly outperformed the classification tree model (0.896 $p = 0.004$). At high sensitivity levels, logistic regression showed the best performance, achieving a specificity of 0.928 at 90% sensitivity, followed by the neural network with one hidden layer (0.916), random forest (0.892), classification tree (0.817), and neural network with two hidden layers (0.782). At a sensitivity of 0.95, logistic regression maintained the highest specificity (0.879). From a clinical perspective, applying a 31% probability cutoff in the logistic regression model could prevent 51% of unnecessary cesarean sections while missing only 5% of acidotic cases.

Discussion

Machine-learning models applied to EFM data show excellent predictive performance for neonatal acidemia. While random forest achieved the highest overall discrimination, logistic regression provided superior performance at clinically relevant high-sensitivity thresholds. These findings suggest that EFM-based predictive models, particularly those focusing on the last deceleration, can serve as effective decision-support tools to reduce unnecessary cesarean deliveries while maintaining patient safety.

[29-P2-24] Dynamic Prediction of Acute Kidney Injury in Heart Failure Patients: Joint Modelling of Serum Creatinine and Time-to-AKI

Presenter: Ziyuan Wang

Authors: Ziyuan Wang [1]; Christiana Charalambous [1]; Timothy Waite [1]; David Jenkins [1]

Affiliations: [1] University of Manchester, Manchester

Introduction

Acute kidney injury (AKI) is a common disease among patients with heart failure (HF). Developing risk prediction models may help clinicians identify high-risk patients and enable timely intervention.

Methods

We developed a dynamic prediction model for time to AKI among patients diagnosed with HF in England, using data from the Clinical Practice Research Datalink. A shared-parameter joint model was applied to link longitudinal serum creatinine trajectories with time to AKI. Model parameters were estimated using Markov chain Monte Carlo. To address the computational burden of model estimation we implemented consensus sampling, which combines partial posterior chains from multiple subsamples of the data. Model performance was evaluated using leave-one-region-out internal-external validation. Predictive performance over time was assessed using discrimination and calibration, and we examined how different choices of observation and prediction windows influenced the performance.

Results

Among 148,689 patients, 38,615 (26.0%) experienced AKI during follow-up. The joint model demonstrated good discrimination and calibration across regions, with time-dependent AUC values of 0.631-0.693, calibration-in-the-large values of 0.746-0.880, calibration slopes of 0.944-1.001, and Brier scores of 0.024-0.094. Model performance varied across different observation windows and prediction windows and was optimal for a prediction window of 1 year, yielding the best balance between discrimination and calibration.

Discussion

Shared-parameter joint model provides a flexible framework for dynamic risk prediction of AKI. We have developed a model that achieves promising performance across a range of prediction windows, and identifying an optimal prediction window may enhance the clinical utility of the model by supporting timely decision-making. The consensus approach offers a practical and scalable solution for developing joint models to large-scale datasets.

P3 - Clinical Trials

Poster area: P3

[29-P3-01] A scoping review on statistical methods and study designs for trials investigating experimental drugs in ultra-rare diseases

Presenter: Elia Gonzato

Authors: Elia Gonzato [1, 2, 3]; Johannes Berkhof [3]; Marjo van der Knaap [1, 4, 2]; Charlie Cao [5]; Rik van Eekelen [6]

Affiliations: [1] Department of Child Neurology, Emma's Children's Hospital, Amsterdam University Medical Center, Vrije Universiteit, Amsterdam; [2] Amsterdam Leukodystrophy Center, Amsterdam Neuroscience, Cellular & Molecular Mechanisms, Amsterdam; [3] Department of Epidemiology and Data Science, Amsterdam UMC location Vrije Universiteit, Amsterdam; [4] Department of Integrative Neurophysiology, Center for Neurogenomics and Cognitive Research, Vrije Universiteit, Amsterdam; [5] Calico Life Sciences LLC, San Francisco; [6] Amsterdam UMC, Amsterdam

Introduction

An ultra-rare disease is a medical condition that affects very few people, with its definition regarding incidence ranging from 1 in 50,000 people to 1 in 100,000. Currently, there are no treatments available for of the majority of ultra-rare diseases, for several reasons. Due to the low number of available patients, setting up a clinical trial is challenging, as is choosing a suitable design. During the last two decades, regulatory bodies have encouraged the development of new trial designs and statistical tools for ultra-rare diseases. (1, 2)

Methods

To synthesise the literature, we conducted a scoping review. We divided our literature search in three pillars: i) projects on improving trial design and statistical tools for rare diseases, including cross-referencing ii) search query on methodological work in databases (i.e. PubMed/Scopus) and iii) search query on clinical trials for (ultra-)rare diseases in clinical trial registries. We also considered guidelines drafted by regulatory bodies regarding drug development for small populations.

Results

We included a total of 227 articles in this review which were divided first into a. randomised or b. non-randomised papers and second into methodological (i and ii) and applied (iii) work. We found three recurring topics: the first were sequential designs that evaluate treatment effectiveness in interim analyses, to stop early for efficacy or futility or otherwise reassess sample sizes. The second was on Bayesian methods that can be used to incorporate information from previous studies. The third was the use of single arm trials in which the comparator was taken from external or historical controls.

We found a gap between methodological and applied work, as the latter mostly consisted in conventional 1:1 randomisation.

Discussion

We will illustrate this gap, elaborate on the reasons of such and discuss two examples of single-arm trials in-depth. (3) We summarize all these advances and provide recommendations for researchers in the field.

References:

1. ASTERIX Available from: https://www.asterix-fp7.eu/.

2. INVENTS Available from: https://invents-he.eu/.

3. Kishnani PS, Corzo D, Nicolino M, Byrne B, Mandel H, Hwu WL, et al. Recombinant human acid [alpha]-glucosidase: major clinical benefits in infantile-onset Pompe disease. Neurology. 200768(2):99-109.

[29-P3-02] Validity and limitations of cross-over designs in degenerative diseases

Presenter: Elena Garcia, Elena Garcia Lara

Authors: Elena Garcia Lara [1]; S. Teerenstra [1]; Kit Roes [1]; Joanna IntHout [1]

Affiliations: [1] Department of Biostatistics, Radboud University Medical Center, Nijmegen

Introduction

Cross-over designs are valuable in clinical research due to their efficiency, particularly in small populations with large between-patient heterogeneity. In rare neurological diseases, however, progressive degeneration may raise concerns about the validity of the treatment effect estimation.

Methods

In this study, we characterise different scenarios based on patient trajectories, treatment effect over time and washout. We also highlight the plausibility of the no-carryover assumption with patients returning to their natural trajectory in the second period regardless of treatment allocation in the first.

Later, we summarise the influence of different profiles on the treatment effect estimation, with common analysis methods based on specific end points or slopes, including linear mixed-effects models. A simulation study compares different approaches with respect to their sensitivity to missed assumptions, missing data, relative efficiency and implications for sample size.

Results

Under the condition of returning to the natural trajectory, even with natural deterioration, mixed-effect models perform similarly as in stable diseases, provided that time is reflected in the model. When that condition cannot be assumed, models that adjust for carryover or models that estimate slopes independently, for example, can be unbiased. These alternatives are often less efficient and each comes with its own limitations. Missing values may reduce the efficiency of the estimations and increase bias.

Discussion

A cross-over design with repeated measurements could be valid in degenerative diseases in certain situations, even if estimation is less efficient with unequal baselines. We provide an evaluation of current approaches in the analysis of cross-over trials in these situations and outline practical recommendations.

[29-P3-03] External Data Strategies in Rare Disease Clinical Trials.

Presenter: Bernard Sebastien

Introduction

Rare diseases present significant challenges for clinical research, particularly in patient recruitment. However, the need for high-quality evidence remains critical, making controlled trials preferable when feasible. A promising approach is to design studies where the control arm (with reduced sample size) is augmented with external or virtual data to increase statistical power. This requires careful attention to potential bias and type I error inflation to maintain regulatory acceptance and scientific rigor. This work compares two strategies: use an external data via matching approach versus simulating virtual data using a model trained on this same external data to complement the control arm.

Methods

Simulations were conducted to evaluate both strategies across multiple scenarios, varying the degree of differences between the RCT population and the external data population, as well as the accuracy of the model used to simulate virtual data (depending on the size of the external database).

Results

The simulations demonstrate that the matching approach generally provides greater statistical power as compared to simulated control arm values. This is because the virtual data approach exhibits higher variability, as uncertainty in parameter estimates is properly accounted for during simulation. Notably, when the external data population differs from the RCT population, the matching approach can result in inflated type I error rates, whereas the virtual data approach remains robust and insensitive to such population differences.

Discussion

While the matching approach can be more powerful, it requires careful implementation, particularly in rare diseases with small treated arms where poorly estimated propensity scores may inflate type I error. The virtual data approach proves more robust across population differences, especially when sufficient external data enables training of accurate predictive models regardless of the RCT treated arm size.

[29-P3-04] A Bayesian Two-Stage Phase II Design with Predictive Probability Monitoring and Capped External-Control Borrowing in IPF

Presenter: Wenjing Xin

Authors: Wenjing Xin [1]

Affiliations: [1] Respiratory & Immunology Biometrics and Statistical Innovation, Astrazeneca, Gothenburg

Introduction

Idiopathic pulmonary fibrosis (IPF) is a progressive, fatal lung disease. Antifibrotic therapies slow decline in forced vital capacity (FVC) but do not halt progression, motivating next-generation agents targeting superiority over existing antifibrotics. Demonstrating superiority is challenging because expected treatment differences are modest relative to high variability in FVC decline, making conventional designs inefficient. An adaptive design leveraging historical antifibrotic control data could enable reliable decisions with fewer patients.

Methods

The trial randomises up to 240 patients 2:1 to the new drug or antifibrotic control. Historical antifibrotic control data are incorporated through a power prior, with the effective number of borrowed patients capped at a limit reflecting how much investigators trust the historical data. At interim, the predictive probability of trial success determines whether the trial stops early for futility or efficacy, or continues to full enrolment. A Bayesian approach is used for all probability models. As benchmark, we evaluate a conventional fixed design with no interim and no borrowing.

Results

Based on simulations, the proposed design controlled incorrect Go decisions below 2% when historical and current control were consistent, and detected the target benefit with 49-70% probability depending on the cap, compared with approximately 40% for the conventional design.

Early futility stopping reduced expected enrolment by 32-35% under the null, requiring 155-162 patients on average versus 240. However, when the true control has drifted from its historical value, aggressive borrowing inflated incorrect Go decisions, whereas a moderate cap limited this inflation.

Discussion

The proposed design shows higher power than a conventional fixed design while requiring fewer patients on average. By enabling earlier Go/No-Go decisions, this approach can accelerate identification of effective new treatments for IPF. We recommend a cap of 40-80 borrowed patients, providing 55-63% power while maintaining robustness to moderate control-rate drift.

This research was funded by AstraZeneca.

[29-P3-05] Adaptive Information Borrowing from Historical Data for Single-Arm Trials Using Density-Ratio-Weighted Learning

Presenter: Yusuke Nakamura

Authors: Yusuke Nakamura [1]; Kota Matsui [2]; Tomohiro Ohigashi [3]; Takafumi Kanamori [4]; Kentaro Sakamaki [5]; Jun Tsuchida [6]; Shigeyuki Matsui [2]

Affiliations: [1] Department of Biostatistics, Nagoya University Graduate School of Medicine, Nagoya; [2] Department of Biostatistics and Data Science, Kyoto University School of Public Health, Kyoto; [3] Tokyo University of Science, Tokyo; [4] The Institute of Science Tokyo, Tokyo; [5] Juntendo University, Chiba; [6] Kyoto Women's University, Kyoto

Introduction

Single-arm trials for new treatment development often lack a control group, necessitating the borrowing of historical trial data to estimate treatment effects. However, differences in covariate distributions between the single-arm trial population and the historical trial population may introduce bias. While existing methods utilize propensity scores or Bayesian modelling, we propose a novel method using density-ratio-weighted learning to adaptively adjust information borrowing for each historical subject under covariate shift assumptions.

Methods

We focus on estimating the average treatment effect, defined as the difference between the average outcome of the target, single-arm trial population where treated and when receiving the control. Under the assumption of covariate shift, where the conditional outcome distribution is shared but covariate distributions differ between target and historical populations (ptarget(x) neq phistorical(x)), our approach employs a two-stage estimation: First, we estimate the density ratio r(x)= ptarget(x)/phistorical(x) to quantify the importance of each historical subject relative to the target population. Second, we estimate the prediction model by minimizing the empirical risk weighted by these density ratios, incorporating L1 regularization. This allows for effective knowledge transfer from the historical population while correcting for distribution mismatch.

Results

Simulation studies demonstrated that the proposed method exhibited stable mean squared error even under large differences in covariate distributions or high-dimensional settings, which inherently exacerbate these differences. Although the estimate variance tended to increase, the estimate mean remained robust compared to conventional methods.

Discussion

The proposed approach allows for effective information borrowing in single-arm trials, mitigating bias caused by covariate shifts while maintaining stability even under high-dimensional covariate settings.

[29-P3-06] Selecting Time Zero in External Control Arms for Single-Arm Trials

Presenter: So Kuroda

Authors: So Kuroda [1]; Eisuke Hida [2]

Affiliations: [1] The University of Osaka, Osaka; [2] The University of Osaka, Suita-city, Osaka

Introduction

Single-arm trials with external controls are increasingly used in rare diseases and other settings where randomized controlled trials are difficult to conduct. Within a target trial emulation framework using external controls, appropriate alignment of time zero, eligibility assessment, and treatment assignment is essential (Hernán et al., 2016). However, in external control data, the same patient may satisfy eligibility criteria at multiple time points, making the choice of time zero nontrivial and potentially influential on bias and precision. Previous studies have investigated how to define time zero in external control settings, but the optimal approach remains unclear. As the definition of the time zero inevitably introduces some degree of bias, our aim is to clarify the statistical properties of the existing methods by simulation studies and to propose an alternative approach to reducing bias.

Methods

We considered disease settings with multiple treatment lines and simulated scenarios in which external control patients could have multiple eligible time points. Time-zero approaches included the first eligible line, last eligible line, random eligible line, all eligible lines, rebalancing approaches, and propensity score-based optimal matching. Their performance was evaluated in terms of bias, precision, and changes in the target population. In addition, these approaches were applied to real-world data to provide descriptive comparisons of treatment effect estimates. We will also explore the development of a refined approach that may further reduce bias relative to the true effect by appropriately combining existing methods.

Results

Simulation and real-world data analyses suggested that different definitions of time zero produced different prior-treatment-line distributions and covariate balance in the external control arm, leading to meaningful differences in bias and precision of treatment effect estimates. Our proposed approach was shown to be practically implementable and to reduce bias compared with existing methods.

Discussion

In externally controlled single-arm trials, the definition of time zero is a key design element that can influence the target population, bias, and treatment effect estimates. Prespecification of time-zero rules and sensitivity analyses across alternative definitions are therefore important.

[29-P3-07] Implicit type I error rate inflation through the use of the one-sample log-rank test

Presenter: Moritz Fabian Danzer

Authors: Moritz Fabian Danzer [1]; Rene Schmidt [1]

Affiliations: [1] Institute of Biostatistics and Clinical Research, University of Münster, Münster

Introduction

Single-arm studies in the early development phases of new treatments are not uncommon in the context of rare diseases or in paediatrics. If an assessment of efficacy is to be made at the end of such a study, the observed endpoints can be compared with reference values that can be derived from historical data. For a time-to-event endpoint, a statistical comparison with a reference curve can be made using the one-sample log-rank test. However, this test does not test for the equality of the actual distributions, but only the conformity with the reference curve. This curve, in turn, is subject to sampling variability.

Methods

Using both analytical considerations and simulations, we examine how the one-sample log-rank test performs under the practically relevant null hypothesis that the distributions in the historical control cohort and the prospective study arm are identical.

Results

For both parametric and non-parametric reference curves, we show that the variance under the null hypothesis of actual interest can be decomposed into the well-known variance of the one-sample log-rank test and a second component that describes the influence of estimation uncertainty in the reference curve. In both cases, this component is mainly driven by the relative size of the single-arm trial with respect to the historical control cohort. In this respect, we also found no advantage to using parametric modelling of the reference curve.

Discussion

Since, in the development of new therapies, it is of interest to assess benefits even at an early stage relative to the actual distribution under the current standard of care - rather than relative to a reference curve, which is also subject to estimation uncertainty - this should be tested using a suitable test. Nevertheless, it is clear that in challenging situations one may have to accept compromises, for example regarding the choice of significance level. Our approaches can also be used to identify the increase in the Type I error rate that one implicitly accepts when choosing a one-sample test.

[29-P3-08] A Population-Adjusted Framework to Quantify Assay Sensitivity in Two-Arm Non-Inferiority Trials

Presenter: Eisuke Hida

Authors: Eisuke Hida [1]; Tomoharu Sato [2]; Satomi Okamura [3]

Affiliations: [1] The University of Osaka, Osaka; [2] Hiroshima City University, Hiroshima; [3] Center for Clinical Sciences, Japan Institute for Health Security, Tokyo

Introduction

In non-inferiority (NI) trials lacking assay sensitivity, ineffective treatments may be erroneously judged non-inferior, resulting in misleading efficacy claims. Regulatory guidelines, including ICH E10 and FDA NI guidance, recommend a three-arm design (test, active control, and placebo) as the gold standard. However, ethical and feasibility concerns often preclude the inclusion of a placebo, limiting the practical implementation of such designs. Consequently, there is a critical need for statistically rigorous methods to quantitatively assess assay sensitivity in 2-arm NI trials.

We propose a population-adjusted method to evaluate assay sensitivity in 2-arm NI trials by leveraging historical active-control versus placebo data and aligning it with the covariate distribution of the NI trial population.

Methods

Assay sensitivity requires demonstrating that the acceptable minimum effective value of the test treatment in a 2-arm NI trial is superior to a placebo (Hida & Tango, 2018). Specifically, the target estimand is the difference between active control and placebo effects in the NI trial population, which is not directly observable. We evaluate this quantity by integrating evidence from the NI trial and external historical trials. To address population heterogeneity, we propose a population-adjustment approach that transports historical treatment effects to the NI trial population. As only aggregate data (AgD) are typically available from historical trials, we formulate the problem as the reverse of a simulated treatment comparison (STC): estimating the historical mean treatment effect under the covariate distribution of the NI trial's individual participant data (IPD). We investigate the performance and practicality of the proposed framework through simulations based on several scenarios using clinical trial data.

Results

Simulation studies and model validation demonstrated that the proposed method provides a robust quantitative evaluation of assay sensitivity across a range of scenarios in NI trials. The proposed framework enables transparent, quantitative, and population-relevant assessment of assay sensitivity in NI trials using historical evidence.

Discussion

Although reliance on historical trials and modelling assumptions may provide a lower level of evidence than the gold-standard design, the framework provides a practical and interpretable basis for decision-making when placebo-controlled trials are infeasible.

[29-P3-09] Flexible and Robust Bayesian Information Borrowing in Clinical Trials with Longitudinal Outcomes

Presenter: David Jesse

Authors: David Jesse [1, 2]; Diane Uschner [3]; Markus Elze [1]; Tim Friede [2, 4]; Francois Mercier [1]

Affiliations: [1] F. Hoffmann-La Roche AG, Basel; [2] Department of Medical Statistics, University Medical Center Göttingen, Göttingen; [3] F. Hoffmann-La Roche, Basel; [4] German Center for Child and Adolescent Health (DZHK), partner site Göttingen, Göttingen

Introduction

Generating robust evidence for new treatments in rare disease and paediatric indications using traditional randomised clinical trial (RCT) designs can be challenging due to the low disease prevalence and hence small sample sizes. To address this problem, a wealth of statistical methods have been developed that allow incorporating information from previous clinical trials into the design and analysis of a new trial (Viele et al. 2014). Bayesian methods are most prevalent in this regard, as reflected in the recent FDA draft guidance on the Use of Bayesian Methodology in Clinical Trials of Drug and Biological Products (Food and Drug Administration 2026).

Beyond the use of external data there are several other options to increase the efficiency of clinical trials.

For instance, whenever possible, it can be advantageous to use a continuous endpoint (as opposed to a derived dichotomous one) and to leverage longitudinal data (Day et al. 2018).

It therefore seems natural to merge these ideas, yet only few of such approaches were considered so far (e.g. Qi et al. 2022). Moreover, the few examples that exist are restrictive in that they assume a specific functional relationship for the longitudinal trajectories, which, if misspecified, can bias the effect estimates and subsequent inferences.

Methods

To address these limitations, we introduce an information borrowing method for incorporating historical

longitudinal control group data into clinical trials.

Our method combines a continuous-time model formulation based on natural cubic splines for the longitudinal trajectories (Donohue et al. 2023) and robust meta-analytic predictive (MAP) priors to incorporate historical data on the control arm (Schmidli et al. 2014).

Results

Through simulations and case studies using data from randomised clinical trials, we demonstrate how our method can increase the efficiency of trial designs while mitigating the consequences of prior-data conflicts.

Discussion

.

[29-P3-10] Bayesian Expert Elicitation for Estimand Construction in Small-Sample Rare Disease: A Cystic Fibrosis Case Study

Presenter: Woosub Shin

Authors: Woosub Shin [1, 2]; Moreno Ursino [1]; Danya Muilwijk [3]; Kors van der Ent [3]; Isabelle Huys [2]; Sarah Zohar [1]

Affiliations: [1] Inserm, Université Paris Cité, Inria, UMR 1346, HeKA, Paris; [2] KU Leuven, Leuven; [3] University Medical Center Utrecht, Utrecht

Introduction

Since its introduction in the ICH E9 (R1) guideline, estimands have been considered an important element in designing clinical trials. While the use of classical and well-known estimands is straightforward, the building of new ones lacks detailed methodological guidance, and this may be attributable to the limited usage of the estimand framework in more complex clinical trials. In our study, we propose a framework for constructing estimands using expert elicitation and Bayesian statistical models.

Methods

We designed a two-stage elicitation process to obtain expert opinions on organoids and on estimand attributes under prespecified scenarios. Using the data, we applied a series of hierarchical joint Bayesian statistical models to identify appropriate estimand attributes that showed the highest relative rankings.

Results

Ten individuals from the ORGESTRA consortium provided opinions, primarily regarding organoids, for designing four clinical trial scenarios during the first-stage elicitation. Next, 39 clinicians, recruited mainly from the European cystic fibrosis networks, contributed expert input on estimand attributes. Across the four scenarios, FEV1, LCI (lung clearance index), and pulmonary exacerbations were selected, either individually or in combinations of two, based on their rankings. Discontinuation of treatment due to lack of improvement was identified as the most likely, and death as the least likely, intercurrent event in all the four scenarios. Full estimand attributes were constructed based on the selected endpoints and intercurrent events.

Discussion

This study demonstrates a framework for establishing clinical trial estimands using Bayesian expert elicitation. With this framework, we intend to support future clinical studies involving novel cystic fibrosis treatments and innovative biological platforms, such as organoids.

[29-P3-11] A critical note on the use of the hypothetical estimand in clinical research

Presenter: Lars Beckmann

Authors: Lars Beckmann [1]; Moritz Felsch [1]; Ulrich Grouven [1]

Affiliations: [1] Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG), Köln

Introduction

The estimand framework ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials was initiated to address the different perspectives of the stakeholders within the health care system, such as patients, physicians, regulatory boards as well as health technology assessment (HTA) bodies.

The framework describes 5 estimands, which differ in the research question but also rely on different data. The treatment policy strategy estimand corresponds to intention-to-treat (ITT) analysis and is recommended for the primary outcome for confirmatory studies for approval as well as for HTAs for all patient-relevant endpoints. The hypothetical estimand, however, considers a counterfactual question under hypothetical assumptions what the value of the endpoint would be, if intercurrent events (ICEs, e.g., disease progression) had not occurred.

In the literature as well as in assessments of medical drugs at the IQWiG (Institute for Quality and Efficiency in Health Care, Germany) the hypothetical estimand is the dominant estimand for endpoints in the categories of morbidity, health-related quality of life and adverse events.

Methods

For the hypothetical estimand, values that could be observed after an ICE may not be collected. And indeed, data collection often ends with an ICE as prespecified in the study protocol. Subsequently, statistical methods such as mixed model for repeated measures as well as Cox proportional hazards models are used to estimate a population-level summary measure for a hypothetical population. The use of this estimand, however, is rarely justified in the spirit of the ICH E9 (R1) addendum.

We will discuss the underlying assumptions and we take a critical look at the practical application of the hypothetical estimand, accompanied by examples.

Results

The hypothetical estimand should be viewed critically with respect to the interpretation and transferability of the results into the health care system.

Discussion

The dominance of the hypothetical estimand does not reflect the intended use. Instead, the estimand is chosen to address data, which are incomplete by protocol. In contrast, data should be regularly collected beyond the ICEs in order to provide the stakeholders with relevant and interpretable information.

[29-P3-12] Per-Protocol Analysis Populations in the Estimand Era: Reframing Adherence-Based Analyses for Decision-Relevant Clinical Trial Interpretation

Presenter: Nicole Close

Authors: Nicole Close [1]

Affiliations: [1] EmpiriStat, Inc., International

Introduction

The ICH E9(R1) estimand framework has shifted clinical trial interpretation away from traditional analysis populations toward clearly defined treatment effects aligned with clinical questions. Despite this shift, per-protocol (PP) analyses remain widely used and are often interpreted as representing treatment efficacy under adherence. In practice, PP populations are commonly defined through post-randomization exclusions based on adherence or protocol deviations, which may introduce selection bias and weaken alignment with the intended estimand. This creates uncertainty about whether PP analyses continue to provide meaningful value in regulatory and clinical decision-making.

Methods

We reviewed implementation of PP and adherence-related analyses across several recent sponsor-led clinical development programs spanning metabolic disease, infectious disease, and medical device trials. For each program, traditional analysis populations were mapped to explicit estimand attributes, including population definition, handling of intercurrent events (e.g., treatment discontinuation, dose modification, rescue therapy), and summary measures. We evaluated whether PP analyses corresponded to identifiable estimand strategies and assessed potential bias pathways related to post-randomization selection and operational variability in protocol deviation capture.

Results

Across programs, conventional PP analyses frequently did not correspond to clearly defined estimands and were sensitive to operational factors such as site monitoring intensity and deviation reporting practices. However, when adherence-related effects were explicitly defined through estimand strategies and supported by aligned estimators and sensitivity analyses, adherence-focused perspectives remained informative for interpretation. These analyses were particularly useful in non-inferiority contexts, exposure-response assessment, and regulatory discussions distinguishing treatment mechanism from real-world effectiveness.

Discussion

Rather than eliminating PP analyses, we propose reframing them as explicitly defined adherence-related estimands with transparent assumptions and operational definitions. This approach maintains interpretability while aligning with regulatory expectations under ICH E9(R1). Greater integration between estimand definition, protocol deviation handling, and SAP implementation is needed to ensure analyses address decision-relevant clinical questions rather than legacy population constructs.

[29-P3-13] Is the use of rescue medication in randomised controlled trials on schizophrenia considered in the statistical analysis or the definition of an estimand as intercurrent event?

Presenter: Julia Frank

Authors: Julia Frank [1]; Anna Hagemeier [1]; Pierce Heiden [1]; Wiebke Müller [1]; Martin Hellmich [2, 1]

Affiliations: [1] Institute of Medical Statistics and Computational Biology, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne; [2] Department of Medical Statistics, University Medical Center Göttingen, Göttingen

Introduction

Rescue medication is often permitted in clinical randomised controlled trials (RCTs). One reason for this is that rescue medication can, for example, prevent patients from dropping out of the trial because the trial drug proves to be ineffective. However, unequal use of rescue medication in the randomised treatment arms may reduce the estimated treatment effect. Therefore, it should be considered in the statistical analysis or at least in the interpretation of results. In particular, the ICH E9(R1) guideline provides a fundamental basis for defining estimands and handling strategies for intercurrent events like rescue medication. This systematic review aims to investigate the proportion of clinical trials on schizophrenia adjusting their statistical analysis to account for the use of rescue medication.

Methods

Phase II and III RCTs of pharmacological treatments in adult patients with schizophrenia were included in this systematic review. Trials published between January 2019 and April 2025 were searched in the databases MEDLINE and PsycINFO and checked for eligibility based on title/abstract and full-text screening. Data extraction was performed by one author using a predefined form and reviewed by a second one.

Results

In total, 1258 articles were checked for eligibility and 37 articles were included after full-text screening. More than two third of the reported trials, permitted rescue medication whereas a few trials did not provide any information on whether rescue medication was permitted. Only a few articles adjusted their statistical analysis for rescue medication or defined an estimand regarding the ICH E9(R1) guideline with rescue medication as intercurrent event. Less than half of all articles that mentioned the use of rescue medication provided details on the extent of such use. No significant differences were observed regarding the start year of the trials.

Discussion

Most investigated clinical trials in schizophrenia permitted rescue medication but did not considered it in their statistical analysis or defined it as intercurrent event. In general, the reporting and sensitivity regarding rescue medication use was insufficient and requires a higher awareness of methodological challenges.

[29-P3-14] Estimation of win-based estimands using a Bayesian framework for randomised controlled trials: a scoping review

Presenter: Chris Selman

Authors: Chris Selman [1]; Melissa Middleton [2]; Leon Di Stefano [2]; Sean Ong [3]; Katherine Lee [2]

Affiliations: [1] Murdoch Children's Research Institute, Parkville; [2] Murdoch Children's Research Institute, Melbourne; [3] Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto

Introduction

Win-based estimands are increasingly used in randomised controlled trials (RCTs) as they prioritise clinically relevant and patient-centred outcomes while providing an intuitive framework for comparing interventions. Bayesian methods offer a flexible framework with probabilistic interpretation of treatment effects and are well-suited to adaptive designs. However, there is limited understanding of how Bayesian approaches can be applied to win-based estimands. This scoping review aims to map current practice, identify gaps, and provide practical examples to support their use.

Methods

We will conduct a scoping review of published and registered RCTs and methodological studies (including unpublished full-text grey literature) that have used Bayesian approaches to estimate a win-based estimand that are available on PubMed, ClinicalTrials.gov, and arXiv, available from 1 January 2012 to 30 April 2026. Screening and data extraction will be conducted systematically, with pilot testing and double extraction to ensure reliability. We will summarise extracted data using descriptive statistics and thematic synthesis for qualitative (free-text) items. Practical examples of identified approaches to estimate win-based estimands using a Bayesian analysis will be illustrated. The review will follow the PRISMA-ScR checklist to ensure rigorous reporting.

Results

This review is ongoing, although preliminary findings suggest a range of approaches for estimating win-based estimands within a Bayesian framework. For time-to-event outcomes, methods include transforming posterior samples from Bayesian survival models into win probability estimates using numerical integration. Hierarchical composite endpoints have been modelled using joint distributions (e.g. bivariate lognormal) within Bayesian futility monitoring trials. For ordinal outcomes, parametric approaches based on Bayesian proportional odds models have been used, with posterior category probabilities being used for win probability and win ratio calculations. Bayesian nonparametric approaches have also been proposed for ordinal outcomes, estimating category probabilities using Dirichlet priors that are then used to estimate a win ratio.

Discussion

This scoping review will provide the first comprehensive overview of how Bayesian methods are used to estimate win-based estimands in RCTs. By synthesising current approaches and identifying gaps in the literature, the review will inform future methodological research, including the evaluation and development of novel Bayesian methods, ultimately contributing to more patient-centred and clinically meaningful inference in RCTs.

[29-P3-15] Application of win statistics as effect measure for composite hierarchical outcomes in randomized controlled trials (WinComp): A systematic review

Presenter: Stefanie von Felten

Authors: Stefanie von Felten [1]; Johanna Ledoux [1]; Rafaela Haltiner [1]; Ulrike Held [1]

Affiliations: [1] University of Zurich, Department of Biostatistics, Zurich

Introduction

The win ratio was introduced by Pocock in 2012 for analyzing hierarchical composite outcomes in randomized clinical trials. Since then the win ratio has gained popularity, particularly in cardiology, and was recently even discussed in one of the JAMA guides to statistics (Cunningham et al, 2026). Its key strength lies in accounting for the differing clinical importance of outcome components through a hierarchical structure. Related methods, including extensions of the Mann-Whitney-Wilcoxon statistic, have been proposed by several authors, though the concept of such "win statistics" dates back earlier, with contributions such as Agresti's generalized odds ratio (1980) and the Finkelstein-Schoenfeld test (1999). The win ratio was criticized because it ignores ties in the inference, and the win odds was suggested as an alternative. Methods to estimate win statistics are sometimes referred to as tournament methods or generalized pairwise comparisons, illustrating the way win statistics are derived.

Methods

We are currently performing a systematic review to systematically identify and describe randomized controlled trials that use win statistics in their design and analysis. Thereby we aim to provide an overview of the use of win statistics across medical fields, highlighting advantages, challenges, and opportunities for assessing overall clinical benefit. The review is registered on PROSPERO (CRD420261299603).

Results

Our literature search identified a total of 2'669 references, which is far more than expected. After de-duplication, the number of references was reduced to 2'595. We are currently screening the references and will be able to present results at the ISCB GMDS 2026 conference.

Discussion

To our knowledge, no comparable systematic review on "win statistics" has been conducted to date.

References:

Pocock S.J., Ariti C.A., Collier T.J., Wang D., The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. European Heart Journal, 33(2): 176-182, https://doi.org/10.1093/eurheartj/ehr352

Alan Agresti, A. (1980) Generalized Odds Ratios for Ordinal Data. Biometrics, 36(1): 59-67. https://doi.org/10.230

Finkelstein, D.M. and Schoenfeld, D.A. (1999) Combining mortality and longitudinal measures in clinical trials. Statistics in Medicine, 18: 1341-1354. https://doi.org/10.1002/(SICI)1097-0258(19990615)18:113.0.CO2
7

Cunningham JM, Fuest S, Burke H. Win Ratio Method for Hierarchical Composite Outcomes in Randomized Clinical Trials. JAMA. 2026335(8):715-716. doi:10.1001/jama.2025.25686

[29-P3-16] Effect of component heterogeneity on statistical power in clinical trials using composite binary endpoints

Presenter: Zerong Chen

Authors: Zerong Chen [1]; Andre Amaral [1]; Maia Lesosky [1]

Affiliations: [1] National Heart and Lung Institute, Imperial College London, London

Introduction

Composite binary endpoints are widely used to increase efficiency in randomised trials, but there is no clear guidance on choosing the best analytical approach. A composite binary endpoint comprises multiple binary outcomes of clinical interest. We conducted a simulation study to compare statistical power and type I error rates to assess the effect of component frequency, effect size and effect direction on established analytical approaches.

Methods

We conducted a simulation study of two-arm individually randomised trials with a composite binary endpoint comprising six components. We conceptualised three levels of heterogeneity (low, medium, high) in component characteristics varying component event rates, treatment effect direction and magnitude, and inter-component correlations. We compared a range of methods, including: collapsed composite analysed by chi-square test and logistic regression count composite analysed by Wilcoxon rank-sum test and Poisson, quasi-Poisson and negative binomial regression win-statistic methods (win ratio, win odds, net benefit) and multivariate models (logistic generalised estimation equations (GEEs) with common, average or distinct effects, and random-intercept logistic generalised linear mixed models (GLMM)). The primary operating characteristics were statistical power and type I error rate over 2,500 replicates per scenario.

Results

Power generally declined as correlation increased, and larger effect sizes of component led to an increase in power. In low-heterogeneity scenarios, the common-effect GEE had the highest statistical power. In medium- and high-heterogeneity scenarios, win-statistics outperformed other methods if components with higher ranks carried large effects, whereas misspecified component ordering reduced power. Methods for count composite, GLMM and common-effect GEE were powerful in most scenarios when all components had treatment effects, while distinct-effect GEE performed substantially better when some components have null effects. The average-effect GEE had lower

power when component effects were unequal, because large magnitude effects in some components were diluted by weaker effects in others. Poisson regression showed marked type I error inflation when correlation among components was present, while GLMM showed mild inflation under strong correlation.

Discussion

No single method was globally optimal. Method choice should be based on the endpoint composition, with our results providing practical guidance. Future methods may focus on win-statistic approaches that are robust to alternative component orderings.

[29-P3-17] Clinical trial evaluation of digital medical devices for clinical decision support: a simulation study

Presenter: Baptiste Archambaud

Authors: Baptiste Archambaud [1]; Sandrine Boulet [1]; Moreno Ursino [1]; Rodolphe Thiebaut [2]; Sarah Zohar [1]

Affiliations: [1] Inserm, Université Paris Cité, Inria, UMR 1346, HeKA, Paris; [2] Université de Bordeaux, Inserm, Bordeaux Population Health Research Center, U1219, Bordeaux

Introduction

Digital medical devices (DMDs) can assist physicians in clinical decision-making by tailoring therapeutic recommendations to individual patient characteristics. Because such DMDs may affect patient outcomes, they should be validated via randomized clinical trials (RCT) and real-world studies. However, their evaluation raises statistical issues, since physicians are free to follow DMD recommendations or not, leading to an observance issue and impacting the observed treatment effect of the DMD. The objective of this study is to propose a simulation based method to assess type 1 error and power of clinical trials evaluating such DMDs.

Methods

We consider the context of a pathology for which several treatments are available. A DMD has been developed to recommend a specific treatment for each patient, based on their personal characteristics. We consider a two-arm cluster RCT, where in control arm, physicians treat patients according to standard practices and in experimental arm, physicians have access to the DMD recommendation to help treatment decision for each patient.

For data simulation, we use a hierarchical framework. Firstly, we define the true model of patients response to treatments, based on mean treatment effects and patient covariates effects. Then, we define the patients response DMD model, which is a biased estimation of the true model. Eventually, physician decision behaviour is built based on physician prior knowledge of the true model and DMD recommendation simulation.

Studied estimand is binary response rate difference between arms in intention-to-treat (ITT) and per-protocol (PP) populations. We compare empirical simulated power against expected calculated power in a traditional two-arm design under comparable settings.

Results

Overall, empirical simulated power aligns with expected calculated power in most simulation scenarios. However, traditional power calculations are overoptimistic compared to empirical values in certain settings, especially with low physician DMD compliance in ITT analysis.

Discussion

This simulation study suggests that traditional sample size calculation might overestimate power when physicians' compliance with DMD is low. Using our simulation method to calculate sample size could give a better estimate when building a trial.

P4 - Causal Methods and Epidemiology

Poster area: P4

[29-P4-01] Should RWD sources be pooled when building an external control group for a single-arm trial? A simulation study

Presenter: Matylde Diouf

Authors: Matylde Diouf [1]; Sylvie Chevret [2]; Lambert Jerome [1]

Affiliations: [1] ECSTRRA Team, IRSL, INSERM UMR1342, Université Paris Cité, Paris; [2] Department of Biostatistics, Hôpital Saint-Louis, AP-HP, Paris

Introduction

In the absence of randomised evidence, unanchored comparisons are increasingly being used to compare the outcomes from a single-arm trial (SAT) with those of external controls derived from real-world data (RWD). However, the choice of the control group directly affects the two key assumptions, underlying the estimation of causal effects: unconfoundedness and overlap. Pooling multiple RWD sources to derive the external control group can present significant challenges. These different sources are either pooled and the differences ignored, or considered separately, regardless of whether the trial objective is to estimate the average treatment effect (ATE) or the treatment effect in the treated (ATT). Whether the heterogeneity in RWD sources justifies pooling them as a control group for the SAT remains an open issue. We aimed to assess whether a multi-source external control group could be pooled, and if so, how to use inverse probability of treatment weighting (IPTW) to provide an unbiased estimate of the ATT on a binary outcome measure.

Methods

We simulated one sample from a SAT, and two external controls arm derived from 2 distinct RWD sources. Patient selection was modelled through a latent/unobserved SAT enrolment eligibility mechanism, with SAT patients sampled only from eligible patients. Exchangeability of study arms was generated by selecting control patients from either the same or a different population. Three scenarios were examined: arm and RWD sources exchangeability arm non-exchangeability with RWD sources (both exchangeable) arm and RWD source non-exchangeability. In each scenario, we compared three propensity score (PS) estimation approaches, either by using: a logistic model with pooled RWD sources a logistic model for each RWD source separately a multinomial model for the data source (generalised PS).

Results

We will present the bias and mean square error of the estimated ATT across the different methods, with varying treatment effects, mean difference in measurable and latent prognostic features across groups, as well as PS overlap across data sources.

Discussion

This study will provide practical guidance on whether RWD sources pooling is appropriate according to target population, analytical method and heterogeneity scenarios.

[29-P4-02] Assessing the Computability of Protocol-Defined Variables in EMR Data: Implications for Trial Emulation

Presenter: Yusuke Sasahara

Authors: Yusuke Sasahara [1]; Hiroaki Moriuchi [2]; Eisuke Hida [1]

Affiliations: [1] The University of Osaka, Osaka; [2] Shionogi & Co., Ltd., Osaka

Introduction

The use of real-world data (RWD), particularly electronic medical records (EMRs), has increased. The target trial emulation framework requires protocol-defined elements, including eligibility criteria and outcomes, to be specified a priori and reproduced using RWD. However, the extent to which these variables can be computable from EMR data is not well characterized and may range from direct computation using structured data to requiring proxy definitions, manual chart review, or being infeasible. Prior studies have primarily examined the availability, completeness, or structure of EMR data elements, whereas the computability of protocol-defined variables has been less systematically evaluated. This study aims to evaluate the computability of eligibility criteria and outcome variables from EMR data using a structured framework.

Methods

We extracted protocol-defined variables, including eligibility criteria and outcome variables, from clinical trial registries. Each variable was classified into five categories within a structured framework to assess computability: (1) directly computable from structured data, (2) computable using proxy definitions, (3) partially computable due to incomplete definitional elements, (4) computable only through manual chart review of unstructured data, and (5) not computable from available EMR data. Distributions were summarized overall and by variable type.

Results

A total of 437 variables were identified across 7 protocols related to thrombolytic agents (eligibility: 329 outcomes: 108). Overall, 21% were directly computable, 11% were computable only with proxy definitions or were partially computable, 62% required manual chart review, and 6% were not computable. Eligibility criteria frequently required manual chart review (68%), particularly for time-window constraints and clinical judgment. For outcome variables, none were directly computable all required proxy definitions or partial reconstruction, most notably for functional endpoints. Direct computability was limited across both categories.

Discussion

These findings indicate that limitations in trial emulation using RWD arise not only from data availability, but also from fundamental constraints in variable computability. The degree of computability varies substantially by variable type, suggesting differential risks of misclassification and bias during variable operationalization. The proposed framework offers a practical tool for evaluating data fitness for use and emphasizes the importance of transparent and systematic variable construction in RWD-based research.

[29-P4-03] Target Trial Emulation to Optimize Trial Setup: Integrating Causal Inference and Machine Learning

Presenter: Charlotte Castel

Authors: Charlotte Castel [1]; Laurent Quinquis [1]; Sophie Swinkels [2]

Affiliations: [1] Danone Global Research & Innovation Center - Medical & Nutritional Science Team, Gif sur Yvette; [2] Danone Global Research & Innovation Center - Medical & Nutritional Science Team, Utrecht

Introduction

Target Trial Emulation (TTE) involves formally specifying the protocol of a hypothetical randomized trial and assessing, within observational data, how far each protocol component - eligibility criteria, exposure definition, covariates, outcome, and analytic strategy - can be reconstructed or approximated. In exploratory or pre-intervention contexts, TTE is a practical tool for strengthening early trial design: it helps determine whether a meaningful and recruitable target population exists in historical data, informs realistic sample size expectations, refines inclusion and exclusion criteria, and supports

hypothesis generation. By quantifying associations under clearly defined analytical choices and by exploring heterogeneity with modern learning methods, TTE also provides early insight into plausible effect size ranges and potential subgroup structures relevant for future interventional trials.

Methods

A target trial protocol was defined, specifying eligibility criteria, exposure strategies, assignment procedures, baseline covariates, and outcomes. Feasibility was assessed by estimating the number of eligible individuals in the observational dataset, informing sample size planning. Exposure was operationalized using distribution-based contrasts (e.g., quartiles) to enhance group separation and reflect contrasts plausibly implementable in a future trial. Confounding was addressed using inverse probability of treatment weighting (IPTW), with weights defined as 1/PS for exposed individuals and 1/(1-PS) for unexposed individuals. Diagnostics included evaluation of propensity score overlap, standardized mean differences before and after weighting, and assessment of extreme weights to ensure adequate covariate balance in the resulting pseudo-population. Weighted linear models described association magnitude and variability and supported derivation of plausible effect size ranges for sample size considerations. Machine learning approaches complemented these analyses: random forests assessed variable importance, while causal forests explored individual-level heterogeneity to identify potential effect modifiers.

Results

The TTE quantified the proportion of historical individuals meeting the target protocol, informing realistic feasibility and recruitment expectations. Weighted models produced consistent association patterns across specifications, and effect estimates supported derivation of plausible effect size ranges for powering prospective trials. Machine-learning models highlighted influential predictors and potential moderators, contributing to refined eligibility criteria and subgroup hypotheses.

Discussion

Although not designed to establish causality in cross-sectional data, TTE offers strong methodological value for early trial setup. By integrating IPTW adjustments, robustness diagnostics, and machine-learning tools for heterogeneity exploration, TTE supports informed decisions on feasibility, population definition, effect size expectations, and subgroup identification, providing a statistically grounded foundation for planning future interventional trials.

[29-P4-04] Causal Effects of Early versus Late Treatment Initiation on Time-to-Event Outcomes Using Observational Data: Protocol for a Methodological Review and Simulation Study

Presenter: Prem Kant Shekhar

Authors: Prem Kant Shekhar [1]; Robin Denz [1]; Nina Timmesfeld [1]

Affiliations: [1] Ruhr-Universität Bochum, Abteilung für Medizinische Informatik, Biometrie und Epidemiologie, Bochum

Introduction

Estimating the causal effect of treatment timing on time-to-event outcomes is a major challenge in observational research. Comparisons of early versus late treatment initiation are particularly vulnerable to confounding by indication, time-dependent confounding, and immortal time bias, especially when treatment can occur only within a restricted follow-up window. This PhD project aims to identify and compare suitable causal inference methods for treatment-timing questions to make their application to real-world clinical data easier. It comprises a methodological literature review and a Monte Carlo simulation study.

Methods

The methodological literature review provides an overview of causal methods for treatment-timing questions in observational time-to-event settings. The simulation study is designed as a neutral comparison of these methods under controlled conditions with known causal structures, with aims, data-generating mechanisms, estimands, methods, and performance measures prespecified according to the ADEMP framework. Monte Carlo standard errors will also be reported to quantify simulation uncertainty.

Results

The protocol follows a target trial emulation framework and specifies one or more causal estimands for comparing early versus late treatment initiation. Candidate analytical approaches include clone-censor-weight methods, marginal structural models with inverse probability weighting, propensity score matching approaches for treatment delay, structural initiation-time models, landmark analysis, and conventional Cox regression. The simulation study design includes data-generating scenarios with baseline and time-varying confounding, informative censoring. Prespecified performance criteria include bias, mean squared error, and computational feasibility.

Discussion

This project compares existing causal methods for treatment-timing analyses in observational time-to-event settings. By combining an overall methodological literature review with simulation-based evaluation, it is intended to support transparent method selection for future empirical analyses.

[29-P4-05] Using electronic health records to evaluate discontinuation of renin-angiotensin system inhibitor treatment in advanced chronic kidney disease in the presence of a grace period

Presenter: Katy Morgan

Authors: Katy Morgan [1]; John Tazare [2]; Patrick Bidulka [1]; Laurie Tomlinson [2]; Leah Pirondini [2]; Ruth Keogh [2]

Affiliations: [1] London School of Hygiene and Tropical Medicine, London; [2] London School of Hygiene & Tropical Medicine, London

Introduction

Renin-angiotensin system inhibitors (RASi) are a key treatment for patients with advanced chronic kidney disease (CKD). However, recent evidence suggests RASi treatment may be associated with accelerated kidney failure with replacement therapy (KFRT), while discontinuation may increase risk of major adverse cardiovascular events (MACE). Electronic health records (EHR) data present the opportunity to investigate the impacts of RASi discontinuation in diverse populations.

Methods

This work uses a longitudinal cohort of adults with advanced CKD from the Clinical Practice Research Datalink (CPRD), a UK-based primary care EHR, linked to Hospital Episode Statistics and death registration data to investigate the causal effect of discontinuing RASi treatment within a grace period of six months versus continuing. Outcomes include all-cause mortality, MACE and KFRT. The target trial emulation framework is used to specify the population, estimands, outcomes and analysis methods. Methodological challenges include time-varying confounding, the presence of a grace period during which treatment may be discontinued, and the irregular observation of individuals over time.

Results

The cohort includes adults with new advanced CKD from 01/01/2007 to 31/03/2022, who take RASi and meet other eligibility criteria. We have identified just under 210,000 patients who may be eligible for our study, with further eligibility criteria to be applied using linked data. The causal contrasts of interest are risk differences at 5 years if the entire study population stopped taking RASi treatment within six months versus if they continued. A large range of potential confounders will be controlled for. We will explore different analysis methods to estimate the effects of treatment strategies with a grace period on time-to-event outcomes, including when there are competing events. These include the clone-censor-weight approach and targeted maximum likelihood estimation incorporating machine learning approaches. The feasibility of applying recently-proposed methods accounting for informative observation of individuals over time will be explored.

Discussion

We will discuss the methodological challenges in addressing questions about treatment discontinuation using EHR data and how we tackled them. In further work, we will explore whether impacts of treatment discontinuation differ according to individual characteristics, and develop prediction under intervention models.

[29-P4-06] Estimating Causal Effects of Dose-Initiation Strategies from Real-World Data

Presenter: Kai Ringwald

Authors: Kai Ringwald [1]; Nina Haug [1]

Affiliations: [1] iOMEDICO, Freiburg im Breisgau

Introduction

Determining appropriate doses for cancer medications is a central problem in oncology, as highlighted by the FDA's Project Optimus initiative. Historically, recommended doses were often based on the maximum tolerated dose. However, many patients may achieve similar outcomes with lower doses, while experiencing fewer toxicities. Estimating the causal effect of post-baseline dose modifications is challenging because such modifications are typically made in response to adverse events, making them subject to time-dependent confounding. In addition, studies that compare "patients with dose reduction" to "patients without dose reduction" often suffer from immortal-time bias if they do not appropriately handle the time at which dose reduction occurs. We therefore focus on a more modest and identifiable estimand: the effect of initial treatment strategy (full starting dose, FSD, vs reduced starting dose, RSD) on outcomes, both marginally and within subgroups defined by the predicted probability of early dose reduction.

Methods

We use inverse-probability-of-treatment weighting to adjust for observed baseline confounding when comparing FSD and RSD, and estimate weighted survival curves for progression-free and overall survival. To characterize heterogeneity of the benefits and risks of RSD, we model the probability of having a dose reduction or toxicity-related treatment discontinuation within the first 6 months, as a function of baseline covariates. Because some patients discontinue or are lost to follow-up before 6 months, we employ inverse-probability-of-censoring weighting to account for censoring due to progression-related treatment discontinuation and other intercurrent events. Patients are stratified into low- and high-risk groups according to their predicted probability of early dose reduction, and the effect of FSD vs RSD is estimated within these subgroups using the weighting approach.

Results

In an illustrative case study of CDK4/6 inhibitor therapy in advanced breast cancer, RSD was not associated with worse progression-free or overall survival in patients at high predicted risk of early dose reduction, while FSD appeared more favourable in patients with low predicted risk.

Discussion

We present a pragmatic framework for evaluating dose-initiation strategies in real-world data using inverse-probability weighting. While unmeasured confounding remains a key limitation, this approach offers a transparent way to investigate dose-optimization questions when randomized trials are infeasible.

[29-P4-07] Discontinuing sodium valproate in young men and women with epilepsy: target trial emulation study

Presenter: Molly Wells

Authors: Molly Wells [1]; Sneha George [2]; Hayley Jones [2]; Georgie Schofield [2]; Iain Buchan [2]; Gregory Lip [2]; Anthony Marson [2]; Glen Martin [1]; Matthew Sperrin [1]; Gashirai Mbizvo [2]

Affiliations: [1] University of Manchester, Manchester; [2] University of Liverpool, Liverpool

Introduction

Randomised controlled trials (RCTs) are the gold standard for evaluating clinical interventions, but are often impractical or unethical. For example, studying the discontinuation of potentially life-saving drugs due to concerns over teratogenicity would require withholding treatment from some patients. This is topical as sodium valproate (VPA) - the most effective treatment for generalised epilepsy - is being restricted due to teratogenicity risks. Target trial emulation (TTE) offers a practical alternative, using causal inference methods to mimic RCTs using observational data. We present a TTE approach for studying VPA discontinuation versus continuation in patients with epilepsy using routinely collected data, with attention to how discontinuation is defined and analysed.

Methods

Routinely collected data from the UK Clinical Practice Datalink (CPRD) and the US TriNetX platform were used to emulate a TTE comparing the effects of discontinuing versus continuing VPA on epilepsy-related death. To account for time-varying exposures and confounding, we used a sequential trial approach, running multiple trials over the study period and averaging the effects. Eligible patients were cloned at trial entry and assigned to both treatment arms simultaneously clones were then artificially censored as they deviated from their assigned treatment, with inverse probability of censoring weights applied to adjust for this censoring and emulate random assignment. Follow-up began at trial entry. Patients were eligible if aged 16-54 years, with one or more epilepsy diagnosis codes, and two or more VPA prescriptions in the previous 12 months.

Results

This presentation will focus on methodological and practical considerations in studying treatment discontinuation using a TTE design with routine healthcare data. This will include defining treatment stopping within prescribing data and applying sequential trials with cloning to address time-varying treatment decisions.

Discussion

This work demonstrates how TTE can be applied to investigate the effects of treatment discontinuation using routine healthcare data. This allows discontinuation to be studied in settings where RCTs are infeasible, supporting safer and more informed decisions about continuing or stopping treatments, such as VPA in patients with epilepsy.

[29-P4-08] Comparing adverse event probabilities in a hypothetical world without consent withdrawals or treatment switches

Presenter: Judith Vilsmeier

Authors: Judith Vilsmeier [1]; Claudia Schmoor [2]; Michael Lübbert [2, 3]; Jan Beyersmann [1]

Affiliations: [1] Institute of Statistics, Ulm University, Ulm; [2] Faculty of Medicine and Medical Center, University of Freiburg, Freiburg; [3] German Cancer Consortium (DKTK) and German Cancer Research Center (DKFZ), Freiburg

Introduction

A question that arises in the analysis of adverse events is how to account for patients who withdraw their consent or switch treatment. Possible approaches are to consider consent withdrawals and treatment switches as competing events or to censor them. However, censoring relies on the assumption that consent withdrawal or treatment switching is unrelated to the treatment or any disease stage, i.e., that censoring is random, which may not hold in practice. These two approaches to handling consent withdrawals and treatment switches in the analysis of adverse events were also discussed by the SAVVY project [1].

Methods

Inverse probability of censoring weighting (IPCW) can be used to evaluate a hypothetical scenario in which treatment switching or consent withdrawal does not occur. For this, patients who remain under observation are up-weighted to represent those who withdraw or switch treatment. We discuss the construction of IPCW estimators in competing events analyses and outline the assumptions required for a causal analysis using data from a randomised study in elderly AML patients investigating the effects of valproate and retinoic acid [2].

Results

All three approaches were used to estimate adverse events probabilities. The results of the approaches were similar, which can be attributed to the small number of patients who switched treatment or withdrew consent.

Discussion

The IPCW approach distinguishes between 'hard' and 'soft' competing events in that hypothetical IPCW calculations are not applied to competing mortality. Since only few treatment switches and consent withdrawals were observed, the weighted pseudo-population created by IPCW does not differ substantially from the observed data.

References

[1] R. Stegherr, J. Beyersmann, V. Jehl, K. Rufibach, F. Leverkus, C. Schmoor, and T. Friede. Survival analysis for AdVerse events with VarYing follow-up times (SAVVY): Rationale and statistical concept of a meta-analytic study. Biometrical Journal, 63:650-670, 2021.

[2] M. Lübbert, O. Grishina, C. Schmoor, et al. Valproate and retinoic acid in combination with decitabine in elderly nonfit patients with acute myeloid leukemia: Results of a multicenter, randomized, 2 × 2, phase II trial. Journal of Clinical Oncology, 38:257-270, 2020.

[29-P4-09] Target Trial Emulation of Adherence Strategies in Nature-Based Social Prescribing for Loneliness: A Causal Inference Analysis of the RECETAS Barcelona Trial

Presenter: Ursula Rochau

Authors: Marjan Arvandi [1]; Sibylle Puntscher [1]; Jill S. Litt [2, 3, 4]; Laura Coll-Planas [5, 6]; Montse Masó-Aguado [5, 6]; Lára R. Hallsson [7]; Ursula Rochau [8]; Uwe Siebert [9, 10, 11]

Affiliations: [1] Institute of Public Health, Medical Decision Making and Health Technology Assessment, Department of Public Health, Health Services Research and Health Technology Assessment, UMIT TIROL - University for Health Sciences and Technology, Hall in Tirol, Austri, Hall in Tirol; [2] Institute for Global Health (ISGlobal), Barcelona Biomedical Research Park (PRBB) Doctor Aiguader, 88 08003 Barcelona, Spain, Barcelona; [3] CIBER Epidemiología y Salud Pública (CIBERESP), Spain, Barcelona; [4] Universitat Pompeu Fabra (UPF), Barcelona, Spain, Barcelona; [5] Research Group On Methodology, Methods, Models and Outcomes of Health and Social Sciences (M3O), Faculty of Health Sciences and Welfare, Centre for Health and Social Care Research (CESS), University o, Barcelona; [6] Institute for Research and Innovation in Life Sciences and Health in Central Catalonia (IRIS-CC), Vic, Spain, Barcelona; [7] Department of Public Health, Health Services Research and Health Technology Assessment, UMIT TIROL - University for Health Sciences and Technology, Hall in Tirol; [8] Institute of Public Health, Medical Decision Making and HTA, Department of Public Health, Health Services Research and HTA, UMIT - University for Health Sciences, Medical Informatics and Technology, Hall in Tirol; [9] UMIT TIROL - University for Health Sciences and Technology, Hall in Tirol; [10] Center for Health Decision Science, Departments of Epidemiology and Health Policy & Management, Harvard Chan School of Public Health, Boston, MA, USA, Boston; [11] Institute for Technology Assessment and Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA, Boston

Introduction

The RECETAS randomized controlled trial (RCT) in Barcelona, Spain, evaluates the Friends in Nature (FiN)-BCN intervention, a nature-based social prescribing program, versus an active control to reduce loneliness. As in many pragmatic RCTs, participants varied in attendance. Because attendance may depend on time-varying characteristics that also affect loneliness, crude adherence analyses may be biased despite randomization. We therefore applied a target trial emulation (TTE) framework to estimate the effect of adherence to NBSP on loneliness at three-month follow-up.

Methods

The NBSP intervention consisted of 9 sessions. Loneliness was measured with the De Jong Gierveld Loneliness Scale (DJGLS 0-11). We emulated per-protocol effects of prespecified adherence strategies using cloning, artificial censoring at strategy deviation, and inverse probability of censoring weighting (IPCW). Two strategy families were considered: dose-based strategies defined by attended sessions (exact and grouped: 0, 1-4, 5-7, and 8-9) and regularity-based strategies defined by attendance patterns over time. Grouped strategies were prespecified as the main analyses, and exact strategies as exploratory analyses. Missing data were handled using multiple imputation by chained equations based on five imputed datasets. IPCW models included baseline covariates, baseline loneliness, prior attendance history, and lagged session-level measures. Weighted generalized estimating equation models estimated mean differences in three-month loneliness scores, and estimates were pooled using Rubin's rules.

Results

No statistically significant differences in three-month loneliness scores were observed. In the grouped dose-based analysis, estimated differences versus 0 sessions were 0.19(95%CI: -0.47to0.84) for 1-4 sessions, -0.02(95%CI: -0.64to0.60) for 5-7 sessions, and -0.05(95%CI: -0.63to0.54) for 8-9 sessions. In the grouped regularity-based analysis, estimates relative to all sessions missed were not statistically significant, although several point estimates were negative. Grouped strategies showed better empirical support and more stable truncated IPCW weights than exact strategies.

Discussion

Although we found no statistically significant evidence that adherence to NBSP affects loneliness at three months, some estimates suggested a tendency towards lower loneliness with higher adherence. These analyses highlight the importance of strategy definition for identifiability and precision. Grouped adherence strategies appeared better supported by the data than exact strategies, suggesting that they may provide a more credible basis for primary causal inference in this setting.

[29-P4-10] Effects of treatment strategies on clinical outcomes and resistance in Enterobacterales infections with chromosomal AmpC: Emulation of target trials

Presenter: Pio Blieske

Authors: Pio Blieske [1]; David Weller [2]; Silvio Brugger [2]; Roger Kouyos [2, 3]; Sebastian Bonhoeffer [1]

Affiliations: [1] Institute of Integrative Biology, ETH Zürich, Zürich; [2] Department of Infectious Diseases and Hospital Epidemiology, University Hospital Zürich, Zürich; [3] Institute of Medical Virology, University of Zürich, Zürich

Introduction

Data from electronic health care records provide a unique opportunity to identify factors that influence antimicrobial resistance and treatment outcomes for individual patients where evidence from randomized clinical trials is limited. However, statistical analysis is often hampered by the heterogeneity of infections, confounding by comorbidities, and ongoing treatment optimisation. Here, we consider the specific case of bacteria in the ESCPM group (Enterobacter cloacae, Klebsiella aerogenes, Serratia marcescens, Citrobacter freundii, Providencia spp., and Morganella morganii), which can switch from low to high resistance by overproducing an initially suppressed chromosomally-encoded AmpC β-lactamase enzyme when certain antibiotics, such as third-generation cephalosporins, are used incorrectly.

Methods

To address confounding and heterogeneity, we use the target trial emulation framework on data collected from electronic medical records of the University Hospital of Zürich, Switzerland. Three treatment arms are evaluated: Carbapenems/cefepime, DNA inhibitors, and inadequate antibiotics

(by AmpC hydrolyzable β-lactams). The outcomes - length of stay, in-hospital mortality, microbiological relapse, and emergence of new phenotypic resistance - are assessed using a Fine-Gray model and logistic regression, respectively. We adjusted for confounding by patient demographics, comorbidities, infection site, and severity (including ICU status, leukocytes, CRP) using inverse probability weighting (IPW) stabilized weights are computed with gradient boosting.

Results

Among 2,869 eligible encounters, DNA inhibitors were rarely administered (5.5%) and required combining treatment arms into adequate versus inadequate therapy. After IPW, covariate balance was achieved (standardized mean differences < 0.1). Using an adjusted Fine-Gray model, we inferred a hazard ratio (HR) for discharge of 0.91 (95% CI: 0.80-1.03) for adequate compared to inadequate treatment, while without adjustment we inferred a HR of 0.89, 95% CI: 0.80-1.00), and for mortality 1.05 (95% CI: 0.75 - 1.47, unadjusted HR 1.11 with 95% CI: 0.82-1.51). For new AmpC-like resistance using weighted logistic regression, we found an adjusted odds ratio (OR) of 1.97 (95% CI: 1.08-3.58) compared to an unadjusted OR of 1.17 (95% CI: 0.74-1.89).

Discussion

This study demonstrates the importance of adjustment for confounding in observational analyses. The findings suggest that while short-term clinical outcomes may be similar, inappropriate use of β-lactams is associated with increased resistance, providing insights for antimicrobial stewardship.

[29-P4-11] A Comparative Study of Tibialis Anterior Tendon Shortening in Combination with Achilles Tendon Lengthening vs Achilles Tendon Lengthening: A Multicentric Target Trial Emulation.

Presenter: Bastian Widmer

Authors: Bastian Widmer [1]; Giusi Moffa [1]; Michèle Widmer [2]; Marys Revaz [3, 4]; Stéphane Armand [3, 4]; Morgan Sangeux [2, 5]; Elke Viehweger [2, 5, 6]

Affiliations: [1] Department Mathematics and Computer Science, University of Basel, Switzerland, Basel; [2] Centre for Motion Analysis, University Children's Hospital Basel (UKBB), Basel, Switzerland, Basel; [3] Kinesiology Laboratory, Geneva University Hospitals and University of Geneva, Geneva, Switzerland, Geneva; [4] Research Center of skeletal Muscle and Movement, Geneva University Hospitals and University of Geneva, Geneva, Switzerland, Geneva; [5] Department of Biomedical Engineering, University of Basel, Basel, Switzerland, Basel; [6] Department of Orthopaedics, University Children's Hospital Basel (UKBB), Basel, Switzerland, Basel

Introduction

Currently, there is no consensus on whether the addition of Tibialis Anterior Tendon Shortening (TATS) provides any advantage over Achilles Tendon Lengthening (TAL) alone in children with cerebral palsy. Therefore, the present study examines whether combining TATS with TAL yields a benefit compared to TAL alone.

Methods

To estimate intervention effects, we follow a formal causal inference framework. To increase transparency and mitigate design-related biases, we applied the principles of target trial emulation in designing an observational study using retrospective data from two clinical gait laboratories. The primary outcome measures were the postoperative gait profile score (GPS) and ankle gait variable score (GVS). To clarify the causal identifiability assumptions and guide the statistical analysis, we constructed causal diagrams informed by literature and expert knowledge. To evaluate the average treatment effect among the treated (ATT) of TAL-TATS relative to TAL, we employed g-methods based on inverse probability of treatment weighting.

Results

A total of 62 individuals undergoing TAL-TATS (14 with bilateral CP) and 21 individuals undergoing TAL (6 with bilateral CP) were included. The estimated effects of TAL-TATS relative to TAL were small and not statistically significant. The ATT for the GPS was -1.48 (95% confidence interval [CI], -7.0 to 4.1), and for the ankle GVS was -1.11 (95% CI, -9.2 to 7.0).

Discussion

The study provides proof of concept for the implementation of target-trial emulation in gait analysis, despite the results being inconclusive. Although point estimates favoured TAL-TATS over TAL, confidence intervals were relatively large, due to the limited sample size, and did not yield statistical significance. Larger studies are needed to determine whether differences between the procedures exist.

[29-P4-12] Difference-in-Differences with multiple time points and right-censoring

Presenter: Emilie Wessel Pedersen

Introduction

In this work, we develop Difference-in-Differences (DiD) methods that accommodate right-censoring under sequential missingness at random. DiD is a widely used causal inference framework for evaluating policy interventions, and recent extensions - most notably by Callaway and Sant'Anna (2021) - generalize classical two-period DiD to multiple time periods and staggered treatment adoption, focusing on group-time average treatment effects and aggregated summaries that capture treatment effect heterogeneity. In applied settings, the choice of time scale is crucial: the interpretation of aggregated estimands depends on the selected time scale, and both identification assumptions and target parameters are inherently time-scale dependent. Many practical applications require scientifically motivated time scales, which often produce unequal follow-up periods across units and induce right-censoring. Our methods address this gap, enabling researchers to use time scales that better reflect the scientific context and underlying confounding structure.

Methods

We formalize identification under a nonparametric model that accounts for sequential missingness at random. Using semiparametric theory, we derive efficient influence functions for DiD estimands under right-censoring and construct corresponding estimators with nuisance estimators based on sequential regression using only observed data and utilizing all of the observed data. These estimators that yield asymptotically valid inference. Our framework naturally extends multi-period and staggered adoption DiD designs while accommodating incomplete follow-up. We also provide guidance on aggregating group-time effects, highlighting how time-scale choices influence both identification and interpretation.

Results

Simulation studies demonstrate that our estimators remain unbiased under varying censoring patterns and time-scale specifications. In applications with staggered treatment adoption, the methods recover interpretable group-time effects and aggregated estimates while properly accounting for right-censoring assuming missingness at random.

Discussion

By integrating right-censoring adjustments into the DiD framework, our work broadens the applicability of causal inference methods to settings with unequal follow-up periods and complex treatment adoption patterns. Researchers can now select time scales aligned with substantive scientific questions without compromising identification or inference validity. These contributions complement and extend the work of Callaway and Sant'Anna (2021) and provide a foundation for future methodological developments in DiD under incomplete follow-up.

References

Callaway, B., & Sant'Anna, P. H. (2021). Difference-in-Differences with multiple time periods. Journal of Econometrics, 225(2), 200-230.

[29-P4-13] Case-Capping versus A-Posteriori Weighting for Over-Representation Adjustment in Multicountry Registries: An ADEMP Simulation Study

Presenter: Anna Sordo

Authors: Gloria Brigiari [1, 2, 3]; Stefania Lando [1, 2, 3, 4]; Ester Rosa [1, 2, 3]; Dario Gregori [1, 2]; Anna Sordo [5]

Affiliations: [1] Unit of Biostatistics, Epidemiology and Public Health, Department of Cardiac, Thoracic, Vascular Sciences and Public Health, University of Padova, Padova; [2] BIOSTAT-X Biostatistics & AI for Biomedical Discovery, Institute of Pediatric Research (IRP) 'Città della Speranza', Padova; [3] PhD Program in Translational Specialistic Medicine "G.B. Morgagni", Curriculum "Biostatistics and Clinical Epidemiology", University of Padova, Padova; [4] Department of Surgery, Oncology and Gastroenterology (DiSCOG), University of Padova, Padova; [5] University of Padova, Padova

Introduction

Voluntary multicountry registries may exhibit over-representation of high-volume centres, potentially inducing representativeness bias and confounding. Case-capping (limiting per-country contributions) and a-posteriori weighting (reweighting to a target distribution) are common remedies, but no systematic simulation comparison has been reported. We address this gap using an ADEMP-structured design (Morris et al., 2019).

Methods

We designed a factorial Monte Carlo experiment varying outcome type (continuous, binary, survival/RMST), overlap severity, dimensionality, treatment-effect heterogeneity, selection mechanism (country-only vs. country-and-covariate dependent), propensity-score misspecification, and sample size. Five methods were compared across 3 pilot scenarios: hard case-capping (=200), design-only weighting, unstabilised IPW (ATE), weight-capped IPW (cap=20), and overlap weights (ATO). Performance metrics - bias, RMSE, 95% CI coverage, type I error (Holm-adjusted), Kish ESS, weight CV, and mean SMD - were computed with Monte Carlo standard errors.

Results

Under the null, case-capping and design-only weighting tended to show considerable bias (up to 0.64) with elevated type I error (74% Holm-adjusted 54%) and limited balance (mean SMD =0.10). IPW reduced bias to 0.009 (continuous), with coverage closer to nominal for binary (98%) and survival (96%) outcomes, though it produced large weights (max =18.7, CV=0.60). Overlap weights showed the lowest bias (0.037), most favourable balance (mean SMD =0.04), bounded weights (max =1.1), highest ESS ratio (0.83), and type I error ~2%.

Discussion

In the scenarios examined, overlap weights compared favourably with case-capping, offering stability without ad-hoc truncation and better balance and ESS retention. Case-capping was associated with elevated type I error, likely because it does not adjust for residual covariate imbalance from differential selection. These results suggest that weighting merits consideration as a preferred strategy, with overlap weights particularly promising when positivity is uncertain. Extensions to doubly robust estimators (AIPW/TMLE) and balancing weights under misspecification are under evaluation.

References: Morris TP et al. Stat Med 201938:2074-2102. Li F et al. JASA 2018113:390-400. Cole SR, Hern'an MA. Am J Epidemiol 2008168:656-664.

[29-P4-14] Association between Choice of Care Setting and Good Death Scale (GDS) Scores: A Generalized Propensity Score Analysis for Multi-Valued Treatments

Presenter: Shota Sonoda

Introduction

The preferred place of care at the end of life (palliative care unit, general hospital, or home) may influence quality of life (QOL), particularly the quality of a "good death," among terminal cancer patients. However, potential confounding by indication - where patients' clinical and social backgrounds influence the choice of care setting - makes it difficult to establish a clear causal link. While the Good Death Inventory (GDI), based on bereaved family surveys, has been widely used, it is subject to recall bias, low response rates, and ethical concerns. In contrast, the Good Death Scale (GDS), evaluated by healthcare professionals based on direct observation, may provide a more objective and practical measure in clinical settings. Because the preferred place of care is influenced by factors such as patients' clinical condition, healthcare resources, and family background, appropriate adjustment for confounding is essential. This study aimed to evaluate the association between the preferred place of care at the end of life and GDS using a generalized propensity score approach.

Methods

We used data from the EASED, a multicenter prospective cohort study including 1,896 terminal cancer patients. The treatment variable was the preferred place of care at the end of life, categorized into three groups: palliative care unit, general hospital, and home. To account for selection bias, we estimated generalized propensity scores (GPS) using a multinomial logistic regression model incorporating 12 pre-specified covariates. Inverse probability weighting (IPW) was employed to estimate the Average Treatment Effect (ATE) across all care settings, adjusting for baseline imbalances. Covariate balance was assessed using standardized mean differences (SMD) for all pairwise comparisons among the three care settings.

Results

Differences in GDS scores across care settings were within approximately one point both before and after IPW adjustment, indicating no clinically meaningful differences.

Discussion

The point estimates remained stable before and after IPW adjustment, suggesting that the observed lack of significant differences in GDS scores was not primarily driven by measured confounding. While IPW improved overall covariate balance, residual imbalance was observed in the general hospital subgroup, likely due to its limited sample size, necessitating cautious interpretation.

[29-P4-15] Low-normal HbA1c and all-cause mortality in adults without diabetes: contrasting flexible parametric survival modelling and targeted maximum likelihood estimation in UK Biobank

Presenter: Safoora Gharibzadeh

Authors: Safoora Gharibzadeh [1]; Zahra Karimi [2]; Cameron Razieh [2]; Clare Gillies [1]; Francesco Zaccardi [2]; Ghanshyam Kacha [3]; Kamlesh Khunti [2]

Affiliations: [1] Leicester Real World Evidence Unit, Diabetes Research Centre, Leicester General Hospital, University of Leicester, Leicester, UK, Leicetser; [2] Leicester Real World Evidence Unit, Diabetes Research Centre, Leicester General Hospital, University of Leicester, Leicester, UK, Leicester; [3] Diabetes Research Centre, University of Leicester, Leicester, UK, Leicester

Introduction

The prognostic significance of low-normal HbA1c in people without diabetes remains uncertain, partly because HbA1c is strongly associated with age, lifestyle, and baseline health. We investigated the association between low-normal HbA1c and all-cause mortality in UK Biobank using complementary time-to-event and causal modelling approaches.

Methods

We analysed 436,072 adults without diabetes at baseline after excluding participants with diagnosed diabetes, diabetes defined by HbA1c, missing HbA1c, HbA1c <20 mmol/mol, and withdrawals. HbA1c was categorised as 20.0-31.9 mmol/mol versus 32.0-47.9 mmol/mol. Confounders were prespecified using a directed acyclic graph. We estimated associations with all-cause mortality using Royston-Parmar flexible parametric survival models on the hazard scale, with sequential adjustment: unadjusted age and sex additionally, ethnicity and the Townsend deprivation index and additionally, smoking, alcohol, body mass index, hypertension, chronic kidney disease, cardiovascular disease, and anaemia. We also estimated marginal risks, risk ratios, and risk differences using targeted maximum likelihood estimation with Super Learner.

Results

Of 436,072 participants, 84,221 (19.3%) were in the low-normal HbA1c group and 351,851 (80.7%) in the mid-range group. The low-normal group was younger, less likely to smoke, and had a more favourable cardiometabolic profile at baseline. In survival analyses, low-normal HbA1c was associated with substantially lower mortality before adjustment (HR 0.64, 95% CI 0.62-0.66), but this association attenuated after age-sex adjustment (HR 0.91, 95% CI 0.88-0.94) and was null in the fully adjusted model (HR 1.01, 95% CI 0.98-1.04). TMLE showed a similar pattern: the unadjusted risk ratio was 0.65 (0.64-0.67), attenuating to 0.91 (0.89-0.94) after age-sex adjustment and to 0.93 (0.91-0.97) after additional adjustment for ethnicity and deprivation.

Discussion

The apparent survival advantage associated with low-normal HbA1c in adults without diabetes was largely explained by confounding. The comparison between flexible parametric survival models and doubly robust causal estimation shows how apparently strong biomarker-mortality associations can disappear after careful adjustment. This illustrates the value of combining prognostic and causal approaches in large observational cohorts.

[29-P4-16] Quantifying the role of stage-at-diagnosis in racial disparities in US cancer survival: a population-based SEER study for 8 common cancers

Presenter: Aisha Vayani

Authors: Aisha Vayani [1]; Mark Rutherford [1]; Sarah Booth [1]

Affiliations: [1] Biostatistics Research Group, Division of Public Health & Epidemiology, School of Medical Sciences, University of Leicester, Leicester

Introduction

Black-White disparities in cancer survival remain substantial in the US. Late stage-at-diagnosis is a plausible and potentially modifiable contributor, but its impact likely varies by cancer site. We quantified, across eight common cancers, how much of the racial gap in five-year all-cause survival (ACS) is mediated by stage-at-diagnosis and translated this into postponable deaths under a hypothetical intervention shifting the stage distribution of Black patients to that of White patients.

Methods

We analysed SEER 21 data for Black and White patients aged 15-89 years diagnosed with malignant solid tumours from 2004-2022 (follow-up to 31 December 2022), covering 41.9% of the US population. Cancers included breast (female), lung and bronchus, colon (excluding appendix), uterine corpus, pancreas, oral cavity and pharynx, rectum and ovary. Stage-at-diagnosis was classified as localised, regional, or distant. A complete-case analysis was used (stage completeness >90%). Mediation was implemented within the relative survival framework using flexible parametric excess-hazard models stratified by sex, applying a period window (2021-2022) to estimate ACS. We partitioned the total causal effect of race on survival into the indirect (stage-mediated) and direct (non-stage pathways) effects, and report the proportion mediated by stage across age groups. We additionally report deaths postponed among Black patients diagnosed in 2022 (inclusive of those with missing stage) [1].

Results

The cohort included 3,054,225 females and 1,389,767 males. Black patients were diagnosed at similar/younger ages and more often with distant-stage disease than White patients, except for rectal cancer in females <75 years. At 5 years, stage-at-diagnosis accounted for roughly half of the disparity for breast (49.87%), colon (48.08% in males 50.69% in females) and corpus uteri (45.33%), and a larger share for lung (70.22% in males 59.85% in females) and oral cavity/pharynx (59.68% in females). Mediation was limited for ovary (18.17%), pancreas (15.41% in males 3.02% in females), oral cavity/pharynx and rectum in males (30.84% and 23.03% respectively). Postponable deaths were time-dependent and greatest where both the indirect effect and case burden were large.

Discussion

Stage-at-diagnosis explains a substantial but heterogeneous share of Black-White survival disparities. Policies promoting earlier diagnosis among Black populations could reduce short-term cancer mortality for several major cancers.

[1]- https://doi.org/10.1002/bimj.201900355

[29-P4-17] Causal Inference with Real World Data in Healthcare Research: A Scoping Review

Presenter: Stella Erdmann

Authors: Stella Erdmann [1]; Leonard Levin [2]; Amina Cheikh Rouhou [1]; Meinhard Kieser [1]

Affiliations: [1] Institute of Medical Biometry, University of Heidelberg, Heidelberg; [2] Francis A. Countway Library of Medicine, Harvard Medical School, Boston

Introduction

Randomized controlled trials (RCTs) are considered the gold standard for evaluating treatment efficacy, as randomization balances confounders and enables causal inference. However, RCTs are often costly, time-consuming, or ethically infeasible. Consequently, other sources for evidence generation are valuable. Real-world data (RWD) - such as electronic health records, medical registries, and health insurance claims - are increasingly used in healthcare research, supported by regulatory initiatives in the United States and Europe, provided that sufficient data quality and methodological rigor are ensured. The growing methodological landscape for causal inference with RWD, however, remains fragmented and inconsistently reported. This study aims to systematically map and structure causal inference methods applied to RWD in healthcare research.

Methods

A scoping review (registered with OSF: https://osf.io/69mcd) is conducted to identify and categorize causal inference methods used in studies based on RWD. The review will examine the frequency and contexts in which these methods are applied, as well as the quality and transparency of their reporting. Particular attention will be paid to the underlying theoretical frameworks, data requirements, and key inferential assumptions. Based on this mapping, a comprehensive taxonomy of causal inference methods will be developed.

Results

PubMed, Embase (Elsevier), and Web of Science were searched for peer-reviewed articles published between January 1, 2010, and January 29, 2026. After deduplication, 2,226 unique records were identified for title and abstract screening. Most studies rely on identification strategies based on selection on observables, typically implemented through regression adjustment or propensity score methods such as matching or weighting, while more advanced causal inference approaches are used only rarely. Preliminary analyses indicate substantial heterogeneity in methodological reporting and justification.

Discussion

This review will provide a structured overview of the evolving methodological landscape for causal inference with RWD in healthcare. The resulting taxonomy is intended to support researchers in selecting appropriate methods, improve transparency and reporting. Ultimately, this work aims to strengthen the validity and interpretability of real-world evidence used in clinical and regulatory decision-making.

P5 - Time-to-event and Longitudinal Methods

Poster area: P5

[29-P5-01] A Functional Fuzzy Entropic Method for Dynamic Survival Prediction via Landmarking

Presenter: Fabrizio Di Mari

Authors: Fabrizio Di Mari [1, 2]; Marco Alfò [3]; Roberto Rocci [3]; Sabrina Giordano [1]

Affiliations: [1] Department of Economics, Statistics, and Finance, University of Calabria, Arcavacata di Rende; [2] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Solna; [3] Department of Statistical Science, Sapienza University of Rome, Rome

Introduction

In clinical settings, the increasing availability of routinely collected longitudinal patient data has made dynamic survival prediction available as an essential tool for improving prognosis. The multi-step landmark approach is a flexible framework that summarizes longitudinal covariates up to a given landmark point and then uses these summaries in survival models. Longitudinal biomarker measurements can naturally be viewed as discrete-time observations of an underlying continuous-time process, which motivates the use of functional data analysis techniques. Within this context, existing approaches use Functional Principal Component Analysis (FPCA) to extract summaries from longitudinal trajectories, which are then used as covariates in the survival model. However, to the best of the authors' knowledge, there is no literature on functional clustering applied to landmark analysis.

Methods

In this work, we propose a novel approach that integrates functional clustering into dynamic survival prediction. We employ a fuzzy entropic extension of the regularized functional reduced k-means algorithm proposed by Gattone and Rocci (2012) to summarize longitudinal trajectories using cluster membership degrees and cluster centroids. These summaries are then incorporated into a Cox model within a landmarking framework for dynamic prediction. The method is extended to take into account non-shared observation times, which represent a fundamental challenge in functional data analysis.

Results

The proposed approach is evaluated on both simulated and real-world benchmark datasets. Preliminary findings suggest that our method provides competitive survival prediction performance compared with existing FPCA-based landmarking approaches in balanced study with common observation times, and it may represent a fundamental step forward in the case of unit-specific observation times.

Discussion

While FPCA provides a low-dimensional representation of functional trajectories, its components are often difficult to interpret, especially when many are required. In contrast, functional clustering yields prototypical trajectories and membership degrees that provide more interpretable summaries of patient biomarker evolution.

Key References:

[1] VAN HOUWELINGEN, H.C. (2007), Dynamic Prediction by Landmarking in Event History Analysis. Scandinavian Journal of Statistics, 34: 70-85.

[3] Gattone, S. A., & Rocci, R. (2012). Clustering Curves on a Reduced Subspace. Journal of Computational and Graphical Statistics, 21(2), 361-379.

Research financially supported by "Patti Territoriali dell'Alta Formazione per le imprese" -CUP: H52C2300009000

[29-P5-02] Comparing joint recurrent event modelling frameworks for dynamic prediction in chronic conditions: an empirical evaluation

Presenter: Thomas Spain

Authors: Thomas Spain [1]

Affiliations: [1] University of Liverpool, Liverpool

Introduction

Joint models linking longitudinal predictors with recurrent event processes offer a principled framework for dynamic risk prediction in chronic conditions. Despite a growing methodological literature, the relative predictive performance of competing joint recurrent event frameworks under varying data conditions remains poorly understood. Practical barriers including convergence difficulties and computational burden further complicate framework selection, yet these are rarely evaluated systematically in applied settings.

Methods

We present an empirical comparison of joint modelling frameworks identified through a systematic review of existing applications in chronic condition research, with three distinct frameworks emerging from current practice: Andersen-Gill (AG), Prentice-Williams-Peterson (PWP), and shared frailty models. Each framework is fitted to multiple longitudinal clinical datasets from chronic condition settings, enabling evaluation across datasets with differing event rates and follow-up structures. A unified validation strategy is applied consistently across all frameworks, with predictive performance assessed using the C-statistic at a fixed time horizon. Convergence rates and computational time are additionally recorded to capture practical alongside statistical considerations.

Results

Results are not yet available at the time of submission. The analysis will characterise how predictive performance, convergence behaviour, and computational demands vary across frameworks and dataset characteristics, providing an evidence base for framework selection in applied recurrent event prediction.

Discussion

This work addresses a practical gap for statisticians developing prediction models that incorporate longitudinal information alongside recurrent event outcomes. By evaluating frameworks across multiple real clinical datasets, this comparison moves beyond theoretical considerations to offer direct, data-driven guidance on when methodological complexity is justified and where simpler approaches remain adequate.

[29-P5-03] Longitudinal Biomarker Kinetics and Survival in ARDS: Insights from the SESAR Trial for Precision Medicine

Presenter: Laurent Renard Triché

Authors: Laurent Renard Triché [1, 2, 3]; Camille Theillière [2]; Bruno Pereira [4]; Sylvie Chevret [5, 3]; Matthieu Jabaudon [1, 2]

Affiliations: [1] Department of Perioperative Medicine, CHU Clermont-Ferrand, Clermont-Ferrand; [2] iGReD, INSERM, CNRS, Université Clermont Auvergne, Clermont-Ferrand; [3] ECSTRRA Team, IRSL, INSERM UMR1342, Université Paris Cité, Paris; [4] Biostatistics Unit, Department of Clinical Research, and Innovation (DRCI), CHU Clermont-Ferrand, Clermont-Ferrand; [5] Department of Biostatistics, Hôpital Saint-Louis, AP-HP, Paris

Introduction

Acute respiratory distress syndrome (ARDS) is a heterogeneous inflammatory lung condition affecting 10-20% of ICU patients, with significant mortality. Recent revisions have increased heterogeneity, complicating the identification of clinically meaningful strategy. Biomarkers may improve prognostic accuracy and treatment response prediction, but their dynamic evolution over time is poorly reported.

This secondary analysis of the SESAR trial (NCT04235608) aimed to explore the relationship between biomarker kinetics and 90-day survival in ARDS to inform precision medicine approaches.

Methods

Within the SESAR trial, which compared two sedative regimens in 687 ARDS patients, we prospectively collected five biomarkers - angiopoietin-2, interleukin-6 (IL-6), IL-8, tumour necrosis factor receptor 1 (TNFR-1), and soluble receptor for advanced glycation end-products (sRAGE) - at six timepoints (from baseline to day 14).

The primary outcome was survival at 90 days. We analysed biomarkers at baseline using a Cox model, and over time using joint modelling and landmark analyses. All models were adjusted for key prognostic variables. Sensitivity analyses were performed to ensure the robustness of the results.

Results

Of the original cohort, 472 patients were included. There were no differences in the baseline or 90-day mortality rate compared to the excluded patients or between the randomisation groups.

Of the five biomarkers, only baseline TNFR-1 was associated with 90-day survival in the multivariable model. By contrast, all biomarkers showed strong associations with survival, that remained consistent across sensitivity analyses. Further landmark analyses confirmed these findings, underscoring the prognostic value of changes in biomarker value beyond single-time point measurements. Furthermore, there was an interaction between the treatment effect and two baseline biomarkers on survival, as well as an interaction between the treatment effect and time on three biomarkers within the longitudinal submodel, reinforcing the clinical relevance of our approach.

Discussion

This study marks the first large-scale, prospective evaluation of biomarker dynamics within an ARDS trial. Integrating kinetic biomarker data with clinical variables, may enable more refined patient stratification and pave the way for biomarker-guided adaptive trials, whether for predictive or prognostic enrichments.

Future work should focus on establishing clinically actionable thresholds and optimal sampling times to implement precision medicine strategies in ARDS management.

[29-P5-04] Joint modelling of longitudinal biomarkers and time to healing in venous leg ulcers

Presenter: Xinyue Zhan

Authors: Xinyue Zhan [1]; Joe Harvey [2, 3]; Jo Dumville [4]; Jamie Sergeant [5]

Affiliations: [1] Centre for Biostatistics, School of Health Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Science Centre, Manchester, UK., Manchester; [2] NIHR Manchester Biomedical Research Centre, Manchester University NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester; [3] Division of Cell Matrix Biology and Regenerative Medicine, School of Biological Sciences, University of Manchester, Manchester; [4] Manchester Centre for Health Psychology, University of Manchester, Manchester; [5] Centre for Biostatistics, School of Health Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Science Centre, Manchester, UK, Manchester

Introduction

Venous leg ulcers (VLUs) are the most common type of chronic leg ulcer, characterised by prolonged healing and substantial physical and financial burden on patients. Identifying prognostic factors based on longitudinal biomarker trajectories could improve prediction of healing and support personalised management. This study aimed to investigate whether longitudinal trajectories of

biomarkers extracted from wound dressings are associated with time to complete healing in patients with VLUs.

Methods

Biomarkers that measured longitudinally were coded B1 to B9 due to protection of intellectual property concerns. The primary outcome was time to complete healing, and the analyses proceeded in several stages. First, biomarker values were transformed to approximate normality. Second, flexible parametric models, Cox proportional hazards models, and accelerated failure time (AFT) models were employed to assess associations between baseline biomarkers and the outcome. Finally, joint models combining linear mixed-effects sub-model with Cox and AFT survival sub-models were applied to investigate associations between longitudinal biomarker trajectories and the outcome. Missing data was handled using joint model multiple imputations, with Blom normal score transformations applied to all biomarkers.

Results

A total of 100 participants were included in the analysis, with a median follow-up time of 21 weeks (interquartile range [IQR]: 14-42 weeks 75% healed within 60 weeks). After adjustment for baseline prognostic factors in the joint models (age, wound area, ulcer duration, infection status, and compression status), B2 and B5 remained statistically significant (adjusted HR [95% CI] for log-transformed biomarkers: B2: 0.528 [0.387-0.720] B5: 0.691 [0.495-0.964]).

Discussion

This study included only 100 patients, primarily older adults (60+) with large, chronic venous leg ulcers, receiving compression therapy and with no infection. The targeted recruitment method means that the sample is not representative of the general venous leg ulcer population but rather reflects a specific subgroup at higher risk for delayed healing. Consequently, the applicability of the observed association estimates may be reduced. Additionally, the modest sample size resulted in relatively wide confidence intervals around estimates. Therefore, these results should be interpreted with caution.

[29-P5-05] Joint Modelling of Longitudinal Clinical and Imaging Data with Survival Outcomes to Characterize Disease Progression

Presenter: Eleni Rosalina Andrinopoulou

Authors: Eleni Rosalina Andrinopoulou [1]; Harry Parr [2]; Aris Perperoglou [2]

Affiliations: [1] GSK, Rotterdam; [2] GSK, London

Introduction

Understanding disease progression is essential for identifying patients at high risk, informing the design of future clinical trials, and supporting strategies aimed at preventing further deterioration. While clinical variables are often evaluated individually, integrating multiple sources of information can provide deeper insight into disease dynamics. In chronic obstructive pulmonary disease (COPD),

longitudinal lung function, together with clinical and imaging markers, may be particularly informative for predicting future exacerbations and guiding prevention-focused interventions. In this work, we aim to improve the identification of high-risk patient phenotypes by integrating multiple data sources using advanced joint modelling techniques.

Methods

We build on an established joint modelling framework that integrates longitudinal and time-to-event data to investigate predictors of time to exacerbations in COPD. Within this framework, associations between key longitudinal outcomes, including lung function and patient-reported outcomes, and exacerbation risk are evaluated under different functional forms. We extend this framework by incorporating longitudinal imaging data through a spatial correlation structure, explicitly modelling dependencies between anatomical lung regions and their contribution to disease progression. This imaging-augmented extension provides additional explanatory value beyond traditional clinical variables, revealing region-specific patterns of lung involvement that are relevant for exacerbation risk.

Results

Initial results demonstrate a relation between exacerbation risk and longitudinal lung function decline, as well as with the underlying value of imaging-derived measures. The joint modelling approach further enabled characterization of patient heterogeneity by linking longitudinal trajectories with individual risk profiles.

Discussion

The proposed joint modelling framework enables a more comprehensive characterization of disease progression, patient trajectories, and risk phenotypes. By jointly leveraging longitudinal clinical, patient-reported, and imaging data, the proposed approach improves identification of patients at high risk of future exacerbations and supports a more integrated understanding of disease dynamics.

[29-P5-06] Joint Modeling of Longitudinal Biomarkers and Breast Cancer Progression for Improving Mammography Screening

Presenter: Evripidis Kapanidis

Authors: Evripidis Kapanidis [1]; Fredrik Strand [2]; Keith Humphreys [1]

Affiliations: [1] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm; [2] Department of Oncology and Pathology, Karolinska Institutet, Stockholm

Introduction

Mammography screening is currently the most common tool utilized for the early detection of breast cancer. Despite its contribution to improving outcomes, false negative and positive examinations persist, exemplifying the need to refine the screening process with additional tools. Artificial intelligence (AI) models for mammography images have been shown to perform comparably to radiologists, and retrospective studies have shown that AI scores increase, on average, in the time period

preceding a diagnosis. Despite these promising results, assessing how AI can be used longitudinally, as an imaging biomarker in clinical practice and how its predictions evolve with respect to tumor progression, remains challenging.

Methods

We develop a framework that jointly models the latent tumor growth, individual screening history, mode of detection (symptomatic or screen detection) and AI scores at each screening examination. Our approach is based on derivations which enable the estimation of all parameters through the use of incident cases data, that is, data including only diagnosed individuals, with a control group of AI scores from images of healthy individuals. Additionally, our framework introduces methodology for distinguishing between healthy individuals and those with latent cancer by incorporating current and prior AI scores. This approach is also applicable to other biomarker data.

We fitted our model to data from a Swedish platform for AI validation in mammography that, for both a large number of controls and cases, includes information on screening histories, AI scores and tumor sizes at diagnosis. We assumed a beta distribution for the AI score distribution dependent on the radiologist decision, latent tumor size and additional covariates.

Results

Overall, estimates of tumor growth and screening sensitivity were consistent with previous studies. Latent tumor size and breast density significantly affected the AI scores. Additionally, using our model, we compared longitudinally incorporating AI scores versus using one AI score at each visit. We found no evidence that incorporation of prior AI scores was associated with improved detection rates.

Discussion

Our model can be used to understand the relationship between diagnostic markers and the natural history of breast cancer, and perform prediction of disease status utilizing full biomarker trajectories.

[29-P5-07] Bland-Altman plot for censored variables with incorporation of information from an additional variable

Presenter: Anne Lotz

Authors: Anne Lotz [1]; Thomas Behrens [1]; Dirk Taeger [1]

Affiliations: [1] IPA - Institut für Prävention und Arbeitsmedizin der Deutschen Gesetzlichen Unfallversicherung, Institut der Ruhr-Universität Bochum, Bochum

Introduction

In Lotz et al. (2025), we demonstrated how Bland-Altman plots can be generated when comparing two measurement methods in the presence of left-censored observations using a multiple-imputation (MI) framework. The present work extends that approach by incorporating information from an additional correlated variable through multiple imputation by chained equations (MICE).

Methods

Using the MICE framework of Laypidus et al. (2014), censored values were imputed with regression models that incorporated the third variable. In each imputed dataset, Bland-Altman statistics (mean difference and limits of agreement) were computed. Rubin's Rules were then used to combine the estimators across imputations. To visualise censoring in the Bland-Altman plot, imputed values for each censored pair were used to compute means and differences, displayed separately from uncensored observations to make censoring patterns explicit.

Results

The method was applied to 176 parallel measurements comparing two welding-fume samplers. Sampler 1 had 15.3% left-censored observations, while Sampler 2 had 11.4% left-censored and 26.7% interval-censored observations. Both sampler measurements were strongly correlated with inhalable manganese concentrations, supporting the inclusion of a third variable in the imputation model. The estimated mean difference between samplers was 0.38, with limits of agreement between -0.56 and 1.31.

Discussion

MICE provides a flexible extension to the classic Bland-Altman approach by accommodating auxiliary variables when handling censored data in method-comparison studies. Applied to welding-fume data, the approach reveals that one sampler systematically measured higher welding-fume concentrations than the other.

References

Lapidus, Nathanael Chevret, Sylvie Resche-Rigon, Matthieu (2014): Assessing assay agreement estimation for multiple left-censored data: a multiple imputation approach. In: Stat Med 33 (30), 5298-5309. DOI: 10.1002/sim.6319.

Lotz, Anne Behrens, Thomas Jöckel, Karl-Heinz Taeger, Dirk (2025): Bland-Altman Plot for Censored Variables. In: Stat Med 44 (13-14), e70147. DOI: 10.1002/sim.70147.

[29-P5-08] Addressing Missing Data When Emulating Hypothetical Interventions in Longitudinal Cohorts: Physical Activity and Mortality in the 1970 British Cohort Study

Presenter: Jamie Wong, Michail Katsoulis

Authors: Jamie Wong [1]; Michail Katsoulis [1]

Affiliations: [1] University College London, London

Introduction

Longitudinal birth cohorts offer the opportunity to emulate hypothetical interventions over the life course. However, analyses of birth cohort data are often affected by missing data. If not handled appropriately, missing data can threaten the validity of estimates, even when applying target trial

emulation. Practical guidance on how to address missing data in this setting remains limited. Using the 1970 British Cohort Study (BCS70) as an example, this study demonstrates an approach to handling missing data when emulating hypothetical physical activity interventions and estimating 21-year mortality risk from age 30 to 51.

Methods

This population-based cohort study used data from BCS70, a British birth cohort of 10,816 individuals born in 1970. We emulated a target trial comparing two hypothetical physical activity interventions from age 30 to 51 years: low physical activity (0 to 3 times per month) and high physical activity (1 to 5 times per week). The estimand was the per-protocol effect of sustained adherence to each intervention on 21-year mortality risk. Missing data were addressed using multiple imputation, with imputation models specified to respect the temporal ordering of variables and include both time-varying covariates and pre-defined predictors of survey non-response, creating 10 imputed datasets.(1) Inverse probability weighting was used to account for adherence, using information from time-fixed and time-dependent confounders. Pooled logistic regression models were used to estimate the standardised risk curves. Non-parametric bootstrapping with 200 samples for each imputed dataset was used to estimate 95% confidence intervals.(2)

Results

3,112 participants were classified under the low physical activity (hypothetical) intervention, and 7,704 under the high physical activity (hypothetical) intervention. The estimated standardised risk of mortality over 21 years was 2.16% (95% CI 1.39%, 3.07%) for the low physical activity group, and 1.31% (95% CI 0.98%, 1.70%) for the high physical activity group. The risk difference between the two hypothetical interventions after 21 years was -0.85% (95% CI -1.79%, 0.01%).

Discussion

Sustained adherence to high physical activity was associated with a moderately lower risk of mortality between ages 30 and 51 years. Methodologically, this study illustrates a practical approach to handling missing data when applying target trial emulation in longitudinal birth cohorts.

[29-P5-09] A graphical framework for the zoo of two-sample metrics: from Kolmogorov-Smirnov paths to Kuiper-type survival processes

Presenter: Lubomir Stepanek

Authors: Lubomir Stepanek [1, 2, 3]

Affiliations: [1] First Faculty of Medicine, Charles University, Prague; [2] Faculty of Informatics and Statistics, Prague University of Economics and Business, Prague; [3] Institute of Computer Science, Czech Academy of Sciences, Prague

Introduction

Graphical path representations of rank-based two-sample procedures provide intuitive insight, but their connections to established nonparametric tests are rarely presented within a unified framework. For unequal group sizes, the pooled rank order induces a directed path whose deviation from a reference trajectory reflects group differences. Related cumulative-process ideas also arise in survival analysis, where censoring replaces the fixed reference line by a risk-set-dependent expected path.

Methods

Several two-sample procedures can be formulated as functionals of cumulative path or score processes generated by ordered pooled observations. In the uncensored setting, maximum deviation, signed range, and label clustering correspond to Kolmogorov-Smirnov-, Kuiper-, and runs-type viewpoints, respectively. This geometric framework extends to right-censored data by representing survival comparisons through risk-set-adjusted score paths related to log-rank and supremum-type tests. Special attention is paid to unequal sample sizes, visualization, and the role of ties.

Results

The framework shows that several seemingly new graphical tests reduce to established statistics in the continuous uncensored setting. It also highlights a methodological gap in survival analysis, where current methods mainly focus on cumulative signed deviation or its supremum. A Kuiper-type range functional may better capture alternatives with opposing early and late effects, such as crossing hazards. The path representation further motivates graphical diagnostics and tie-aware computational developments.

Discussion

The main contribution is a unified process-based perspective on two-sample inference across classical rank tests and censored survival settings. This perspective helps distinguish rediscoveries from genuine opportunities for methodological development. A promising next step is the formal study and calibration of a Kuiper-type survival statistic, together with graphical tools for teaching, diagnostics, and comparison of metric choice.

[29-P5-10] A Rank-Based Test for Dependent Censoring in Survival Analysis

Presenter: Stylianos Tzortzakis

Authors: Stylianos Tzortzakis [1]

Affiliations: [1] National and Kapodistrian University of Athens (NKUA), Athens

Introduction

In survival analysis, the assumption of independent censoring is fundamental for the unbiased estimation of net quantities, such as the survival and hazard functions. However, when data exhibit

dependent censoring (Not Censored At Random - NCAR), standard estimates become biased, leading to non-identifiability issues. Understanding whether our data involve dependent censoring cases can help us decide if the causal independence assumption is applicable. This research aims to provide a diagnostic test to indicate whether the survival data at hand dictate dependent censoring.

Methods

The proposed methodology operates under the assumption of proportional crude hazards, akin to the Koziol-Green model. By utilizing Sklar's theorem and the framework of copulas, we analyze the probability of observing an event given the minimum time X, denoted as $P(\delta=1 X=x)=p(X)$. We mathematically prove that a constant probability of this kind denoted as p, leads to independence between the time-to-event and censoring mechanisms. Based on this theoretical foundation, we develop a Wilcoxon-type rank sum test adapted for survival data, utilizing extended ranks for both events and censored observations.

Results

We demonstrate that under the null hypothesis (H0) of constant probability p, which implies independent censoring, the actual observed ranks for both events and censored observations follow a discrete uniform distribution. The proposed test evaluates deviations from this discrete uniformity. Specifically, the critical region targets scenarios where censored individuals are disproportionately observed at early or late stages of a study (i.e., extreme rank sums), which is a strong indicator of dependent censoring. For large sample sizes, the sum of these ranks follows the normal distribution.

Discussion

The proposed rank-based test offers a practical, distribution-free approach to verify the independence assumption in survival mechanisms. By identifying dependent censoring through directly estimated crude quantities and ranks, researchers can make informed decisions regarding the validity of standard survival models and the necessity for restrictive assumptions or correction factors.

References

Heitjan, D. F., & Rubin, D. B. (1991). Ignorability and coarse data. The Annals of Statistics, 19(4), 2244–2253.

Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. Proceedings of the National Academy of Sciences, 72(1), 20–22.

Koziol, J. A., & Green, S. B. (1976). A Cramer-von Mises statistic for randomly censored data. Biometrika, 63(3), 465–474.

[29-P5-11] When the going gets tough: A comparison of time-to-event modelling methods when hazards are not proportional

Presenter: Kristy Robledo

Authors: Kristy Robledo [1]; Ian Marschner [1]; Gillian Heller [1]

Affiliations: [1] NHMRC Clinical Trials Centre, University of Sydney, Camperdown

Introduction

Randomised clinical trials (RCTs) routinely assess time-to-event outcomes, such as mortality, using Cox regression models. However, these models rely on the proportional hazards (PH) assumption, which is frequently violated in RCTs with long-term follow-up, particularly when RCT results show large treatment benefits. Using the Long-term Intervention with Pravastatin in Ischaemic Disease (LIPID) trial as a motivating example, we compare a range of semi-parametric, parametric, and non-parametric approaches for estimating treatment effects under non-proportional hazards, framed within the estimand framework.

Methods

The LIPID trial randomised 9014 participants to pravastatin or placebo and followed them for 16 years ~85% of placebo participants initiated statin therapy after the blinded RCT ended at 6 years, generating substantial non-proportional hazards.

We compare 5 methods: semi-parametric (Cox, time-partitioned Cox), parametric (accelerated failure time [AFT], generalised additive models for the location, shape and scale [GAMLSS]), and non-parametric (quantile regression), and interpret treatment effects within the estimand framework.

Results

We found that standard Cox regression suggests a modest benefit of pravastatin (HR 0.93, 95%CI:0.88-0.99) however the PH assumption is violated. Time-partitioned models reveal a stronger effect during the RCT phase (HR 0.79, 95%CI:0.70, 0.89) and no effect post-RCT (HR 1.00, 95%CI:0.93-1.07). AFT models found a 4.7% (95%CI:0.7-8.8) increase in median survival for pravastatin. GAMLSS revealed effects of pravastatin on both median survival and variability. Median survival increased by 19.1% (95%CI:17.5-20.8) with pravastatin, while variability decreased by 42.8% (95%CI:45.2-40.5).

Discussion

This comprehensive comparison highlights the importance of selecting appropriate models aligned with the estimand of interest, especially in the presence of non-proportional hazards. The findings underscore the value of flexible modelling strategies aligned with clinically meaningful estimands. Different modelling strategies illuminate different aspects of the treatment effect: absolute vs. relative differences, early vs. late effects, and distributional changes. Incorporating estimands ensures transparent interpretation and may improve decision-making in long-term trials where treatment switching or delayed effects are common.

[29-P5-12] Modeling multiple time scales in survival analysis with applications to smoking-related mortality

Presenter: Makiko Mieno

Authors: Makiko Mieno [1]; Noriko Tanaka [2]

Affiliations: [1] Dept. Medical Statistics, Center for Information, Jichi Medical University, Shimotsuke; [2] Dept. Biometry and Genetics, Institute of Biomedical Sciences, Fukushima Medical University, Fukushima

Introduction

Modeling event hazards in survival analysis often requires consideration of multiple time scales. We consider settings in which a secondary time scale affects the primary time to an event, such that event timing depends on when the secondary process begins. For example, life expectancy may depend on both the timing of smoking initiation and smoking duration. Since Lexis' proposal in 1875, various methods have been developed to address multiple time scales, including approaches focusing on time-scale choice, the role of entry age, and flexible hazard modeling. However, no existing method fully captures this problem. In this study, we compare existing methods using observational data and evaluate their potential bias through simulation.

Methods

We analyzed data from a prospective autopsy cohort conducted at a geriatric hospital in Japan, which included 2,721 autopsy cases with smoking history between 1995 and 2019 (mean age at death 79). Four strategies were evaluated: (1) Cox regression with attained age as the primary time scale, incorporating other time scales as covariates (2) Oakes's two-time-scale model (3) a multistate (semi-Markov) model and (4) a flexible rate model using splines.

Results

The hazard function varied over age of smoking initiation and was not adequately summarized by a constant hazard ratio. Among individuals initiating smoking before age 50, the association between smoking duration and mortality was consistent with conventional findings, whereas among those initiating at age 50 or older, longer smoking duration was associated with longer survival. In this elderly cohort, Cox models including smoking duration and age of smoking initiation as a covariate tended to overestimate the effects of smoking (HR: 8.6 95% CI: 6.33-11.7) and smoking duration (HR: 0.79 95%CI: 0.761-0.825). The multistate model gave most plausible estimates for average smoking effect (HR: 1.56 95% CI: 1.37-1.77) and age of smoking initiation (HR:0.54 95% CI: 0.421-0.695).

Discussion

Models that do not account for the interplay between smoking initiation timing and duration may yield biased or misleading estimates. Flexible modeling approaches, including multistate and flexible hazard models, may offer advantages over conventional approaches in capturing these complex relationships, although further evaluation is needed.

[29-P5-13] Non-Linear Power: A Shape-Based Framework for Sample Size in Survival Analysis with Splines

Presenter: Stefania Lando

Authors: Stefania Lando [1, 2, 3, 4]; Ester Rosa [1, 3, 4]; Gloria Brigiari [5, 3, 4]; Dario Gregori [5, 3]

Affiliations: [1] Unit of Biostatistics, Epidemiology and Public Health, Department of Cardiac, Thoracic, Vascular Sciences and Public Health, University of Padova, Padova, Padova; [2] Department of Surgery, Oncology and Gastroenterology (DiSCOG), University of Padova, Padova, Italy, Padova; [3] BIOSTAT-X Biostatistics & AI for Biomedical Discovery, Institute of Pediatric Research (IRP) 'Città della Speranza', Padova, Italy, Padova; [4] PhD Program in Translational Specialistic Medicine "G.B. Morgagni", Curriculum "Biostatistics and Clinical Epidemiology", University of Padova, Padova, Padova; [5] Unit of Biostatistics, Epidemiology and Public Health, Department of Cardiac, Thoracic, Vascular Sciences and Public Health, University of Padova, Padova

Introduction

Flexible modeling of continuous exposures in Cox regression is often achieved using restricted cubic splines (RCS), which allow non-linear effects while retaining interpretability. However, standard sample size calculations typically assume linear (1-degree-of-freedom) effects, creating a mismatch when multi-parameter spline terms are used in analysis. This can lead to underpowered tests for non-linearity and unstable hazard ratio estimates. We propose a framework to align study design with intended non-linear analyses.

Methods

We introduce Non-Linear Power (NIP), a design-by-shape approach that specifies a clinically meaningful hazard ratio curve and maps it to spline coefficients. Power is derived using a non-central χ^2 approximation, where the non-centrality parameter depends on the expected number of events and the per-event Fisher information. To complement power-based planning, we incorporate a stability criterion based on global shrinkage, quantifying the additional "parameter tax" associated with increasing spline degrees of freedom. The minimum required events are defined as the maximum between power- and stability-based requirements.

Results

In an illustrative scenario (anticipated event fraction 0.10, R^2 = 0.15, shrinkage target 0.9), moving from a linear effect to RCS models with 3 and 4 knots approximately doubles and triples the number of events required for stable estimation. For curvature detection, required events depend on the strength of the non-linear signal for example, moderate curvature implies approximately 52-73 events for tests with 1-3 degrees of freedom. Depending on assumptions, either power or stability may dominate the design constraint.

Discussion

Designing studies under linear assumptions when non-linear effects are intended can result in inadequate power and unstable estimates. The proposed NIP framework enables power calculations aligned with the shape of the hypothesized effect, while the parameter tax quantifies the cost of model flexibility. Together, these tools support more coherent sample size planning for non-linear Cox regression analyses.

[29-P5-14] Improving the selection and accuracy of survival models: a bayesian spike and slab approach.

Presenter: Joanna Pautonnier

Authors: Joanna Pautonnier [1, 2]; Pascal Roy [3, 2, 1]; David Causeur [4, 5]

Affiliations: [1] Université Claude Bernard Lyon 1, Lyon; [2] LABORATOIRE DE BIOMÉTRIE ET BIOLOGIE ÉVOLUTIVE, Lyon; [3] Hospices civils de Lyon, Lyon; [4] Institut de recherche mathématique de Rennes, Rennes; [5] L'Institut Agro, Rennes

Introduction

In the era of precision medicine, survival models are essential for predicting individual risk and guiding clinical decision-making. To improve predictive accuracy, modern approaches increasingly incorporate flexible functional forms and interactions. However, this added complexity raises challenges in variable selection and uncertainty quantification. Classical methods are often computationally expensive and provide unreliable confidence intervals in high-dimensional or highly flexible settings. In addition, widely used models such as the Cox proportional hazards model rely on restrictive assumptions, particularly the proportionality of hazards, and do not directly estimate the baseline hazard (Cox, 1972).

Methods

We propose a Bayesian framework for flexible hazard modeling based on an additive predictor that allows for non-linear and time-dependent covariate effects using penalized splines. Unlike existing approaches that approximate the hazard likelihood via piecewise Poisson models, our method directly relies on the hazard likelihood, thereby reducing computational burden and avoiding data expansion. Variable and function selection are achieved through a spike-and-slab prior (peNMIG), applied at the level of coefficient groups, enabling simultaneous selection, shrinkage, and interpretation within a hierarchical model (Scheipl, 2011). Numerical integration techniques are used to approximate the cumulative hazard.

Results

A simulation study was conducted to evaluate the proposed method in terms of variable selection performance, estimation accuracy (bias and mean squared error), and coverage probabilities of credible intervals. The results show that the approach effectively identifies relevant covariates and functional effects while excluding irrelevant ones. Compared to existing methods, it provides improved coverage probabilities, indicating more reliable uncertainty quantification, while maintaining low bias and small mean squared errors.

Discussion

The proposed approach offers a unified and flexible Bayesian framework for survival analysis that addresses key limitations of existing methods. By avoiding the Poisson approximation and relaxing the proportional hazards assumption, it enables a more accurate representation of complex risk dynamics over time. Furthermore, it ensures coherent uncertainty quantification and interpretable

model selection. This makes it a promising tool for applications in precision medicine, where reliable individual predictions and uncertainty assessment are crucial.

[29-P5-15] A non-parametric Bayesian approach for the one-sample log-rank test accounting for reference curve sampling variability

Presenter: Chloé Szurewsky

Authors: Chloé Szurewsky [1]; Gwénaël Le Teuff [1, 2]

Affiliations: [1] CESP, INSERM U1018, Université Paris-Saclay, Villejuif; [2] Biostatistics and Epidemiology Office, Gustave Roussy, Université Paris-Saclay, Villejuif

Introduction

The one-sample log-rank test (OSLRT) under the proportional hazards (PH) assumption or its Bayesian counterpart are commonly used in oncology to compare the survival curve of an experimental group with an external control survival curve S0(t). They are particularly useful when designing a randomised clinical trial is not possible. They require both (i) the specification of a parametric distribution for S0(t) to calculate the expected number of events E, and (ii) assume no sampling variability of the external control group.

Methods

We develop a new flexible approach based on a non-parametric Bayesian modelling of S0(t) (indirectly the cumulative hazard function Lambda0(t) to estimate E = sumi =1n Lambda0(.)) using a Mixture of Dirichlet Process, allowing to circumvent these 2 restrictive assumptions. A Dirichlet process prior is put on the cumulative distribution, F0(t), with an exponential base distribution and a Gamma prior on its rate parameter. The performances are compared to the frequentist and Bayesian OSLRT by simulations of single-arm trials under PH scenarios. The survival times of the experimental group are generated with an exponential distribution, and those of the external control with an exponential distribution (100 patients with 5% of censoring and lambda = 0.35). The simulation parameters are the sample size of the experimental group, the exponential censoring rate, and the hazard ratio. A single-arm trial in children with neuroblastoma evaluating an inhibitor is used for illustration.

Results

The Bayesian non-parametric OSLRT is less conservative than the OSLRT, mOSLRT and its Bayesian version, in particular for large sample sizes. The non-parametric Bayesian OSLRT is more powerful than the OSLRT for small sample sizes and has the same power for larger sample sizes (more than 100 patients). These results are consistent with those reported in the literature when the sampling variability is accounted for in the frequentist OSLRT.

Discussion

We propose a new assumption-free method based on a non-parametric Bayesian process to compare one survival curve to a reference curve. Further work is to adapt this process to non-PH, to incorporate information from historical control data and to investigate direct non-parametric Bayesian

estimation of the cumulative hazard function of the external control group with a Gamma or Beta process.

[29-P5-16] Defining win statistics for censored data

Presenter: Andreas Kühnapfel

Authors: Andreas Kühnapfel [1]; David Petroff [1]; Miroslav Bacak [1]

Affiliations: [1] Universität Leipzig, Leipzig

Introduction

The so-called win statistics have attracted attention in recent years as an alternative approach for dealing with composite endpoints. They use a non-weighted hierarchy in the components to define a "win", "loss" or "tie" when comparing the outcome of two patients [1]. In particular in the time-to-event setting, there may be a large amount of censored information and no accepted approach to dealing with it.

Methods

We propose an approach that estimates the probability for a "win", "loss" or "tie" for pairs of patients with censorship. Simulations allow us to compare the properties of our approach to those of the naive approaches used until now.

Results

In many realistic scenarios, the differences are quite large between methods and the newly proposed probabilistic approach greatly reduces the number of strict ties. Research on simulations is ongoing.

Discussion

Until now, censored data have not been handled properly within win statistics. Especially in trials with administrative censoring, this can have a very large impact on results.

References

[1] Pocock SJ, Ariti CA, Collier TJ, Wang D. The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. Eur Heart J 2012 33: 176-82. https://doi.org/10.1093/eurheartj/ehr352

[29-P5-17] Modelling heterogeneous treatment effects in time-to-event data: a comparison of hierarchical Lasso-regularized Cox models and causal survival forests

Presenter: Sophia van Goor

Authors: Sophia van Goor [1]; Marta Fiocco [1]; Hans Merks [2]; Marta Spreafico [1]

Affiliations: [1] Mathematical Institute, Leiden University, Leiden; [2] Princess Máxima Center for Pediatric Oncology, Utrecht

Introduction

Randomized controlled trials represent the benchmark for evaluating treatment efficacy, typically reporting average treatment effects across all participants. However, patients may respond differently due to individual characteristics, leading to heterogeneous treatment effects (HTEs). Standard approaches to explore HTEs, such as subgroup analyses or likelihood ratio tests for pre-specified treatment-covariate interactions, may miss complex patterns of heterogeneity. More flexible methods, including regularized regression techniques and machine learning approaches, can model complex interactions and better identify patient subgroups with different treatment benefit.

Methods

In this study, we compared hierarchical lasso-regularized Cox model (HL-Cox) and causal survival forests (CSF) for estimating HTEs on survival outcomes. HTE was expressed as the conditional average treatment effect (CATE), defined as the difference in restricted mean survival time (RMST) between alternative treatments. We first evaluated the performance of these methods in a simulation study and then applied them in a retrospective analysis of 640 patients from the Euro Ewing 2012 (EE2012) trial, in which primary subgroup analyses had found no significant HTEs between the two chemotherapy regimens for Ewing sarcoma, VIDE and VDC/IE.

Results

Both methods provided evidence of heterogeneity, highlighting the value of flexible approaches for detecting HTEs. In the EE2012 trial, the HL-Cox approach estimated 5-year CATEs ranging from 17 to 412 days of event-free survival gain under VDC/IE versus VIDE, with local extension towards other sites and regional lymph node involvement driving heterogeneity. CSF yielded 5-year CATEs ranging from -52 to 237 days, with BMI, kidney function, ejection fraction and age emerging as the main heterogeneity contributors. Overall, the findings suggest that, together, the HL-Cox and CSF approaches offer a more complete picture, offering complementary insights into which patients gain the greatest benefit from VDC/IE versus VIDE.

Discussion

Both methods consistently identify HTEs, although their estimates differ due to different modelling assumptions. The two methods are complementary rather than competitors: together they offer a more flexible way to identify HTEs.

[29-P5-18] An Integrated RAEN-CooPeR Framework for High-Dimensional Variable Selection in Competing Risks Survival Models

Presenter: Md. Monowar Hossain

Authors: Md. Monowar Hossain [1]; Md Hasinur Rahaman Khan [2]

Affiliations: [1] Institute of Statistical Research and Training (ISRT), University of Dhaka, Narayanganj; [2] Institute of Statistical Research and Training (ISRT), University of Dhaka, Dhaka

Introduction

High-dimensional survival analysis under competing risks remains a fundamental challenge in modern biomedical research, particularly in genomic applications where complex correlation structures and rare event types prevail. Conventional penalized approaches such as LASSO often exhibit instability in the presence of strong multicollinearity, while standard cause-specific modeling strategies fail to leverage shared covariate information, thereby reducing efficiency. To address these limitations, we propose a novel integrated framework - Random Approximate Elastic Net-Cooperative Penalized Regression (RAEN-CooPeR) - designed to achieve stable variable selection and enhanced predictive performance in high-dimensional competing risks settings.

Methods

The proposed framework consists of two synergistic phases. In Phase I, RAEN employs ensemble-based prescreening through decorrelation-driven partitioning of predictors into weakly dependent subsets. Each subset is analyzed via unpenalized proportional subdistribution hazards models, followed by weighted least squares approximation and bootstrap aggregation to yield a stable reduced feature space. In Phase II, CooPeR applies an iterative, feature-weighted elastic net penalty that facilitates information sharing across cause-specific subdistribution hazards models, thereby reinforcing common signal structures without requiring predefined grouping assumptions. To enhance interpretability, we further integrate a compact neural architecture with residual connections and cause-specific attention mechanisms, enabling intrinsic and transparent feature attribution.

Results

Extensive simulation studies are anticipated to demonstrate that RAEN-CooPeR consistently outperforms state-of-the-art methods - including penalized Cox models, CoxBoost, random survival forests, and DeepHit - in terms of variable selection accuracy (sensitivity, specificity, and false discovery control) and predictive performance (scaled mean squared error). Application to bladder cancer genomic data (GSE5479 n=385, p=1,381) reveals biologically plausible markers associated with multiple competing outcomes, with superior discrimination (concordance index) and calibration (integrated Brier score).

Discussion

The proposed RAEN-CooPeR framework effectively unifies high-dimensional screening with cooperative penalization, addressing both correlation-induced instability and information fragmentation across competing risks. Unlike conventional black-box learning approaches, it offers a principled balance between predictive accuracy and interpretability - an essential requirement in clinical decision-making contexts.

P6 - Evidence and Decision Sciences

Poster area: P6

[29-P6-01] Statistical methods for extrapolation between populations: a review

Presenter: Lingjiao Wang

Authors: Lingjiao Wang [1]; Francois Mercier [2]; Astrid Jullion [3]; Xuan Zhu [3]; Christian Röver [4]; Marc Vandemeulebroecke [5, 6]; Martin Posch [7]; Sarah Zohar [8]; Nigel Stallard [1]

Affiliations: [1] University of Warwick, Coventry; [2] F. Hoffmann - La Roche Ltd, Basel; [3] Novartis Pharma AG, Basel; [4] Universitätsmedizin Göttingen, Göttingen; [5] Radboud University Medical Center, Nijmegen; [6] Bayer BCC AG, Basel; [7] Medical University of Vienna, Vienna; [8] Inserm, Université Paris Cité, Inria, UMR 1346, HeKA, Paris

Introduction

Extrapolation refers to statistical approaches that support inference on treatment effects in a target population using data from completed clinical trials conducted in a source population. This is particularly important when only limited data are available in the target population. Extrapolation may be based on multiple sources of information, including pharmacokinetic, pharmacodynamic, mechanistic, or clinical outcome data. This review focuses primarily on outcome-based approaches, particularly methods that enable borrowing of information across populations.

Methods

We conducted a review of statistical methodologies for extrapolation, focusing on approaches that incorporate outcome data from source populations to make inference in target populations. Methods were identified across a range of settings, including extrapolation across age groups and across different clinical indications.

Results

The review identifies 218 extrapolation approaches from 159 papers, which can be broadly classified into three main categories: Bayesian, frequentist, and causal inference methods. These methods vary in their assumptions, implementation, and applicability across different extrapolation scenarios.

Discussion

This review provides a comprehensive and up-to-date overview of statistical methods for extrapolation, offering a foundation for further methodological development and applied research. It also identifies existing gaps in the literature and thus provides a basis for future work on the development of new extrapolation methods and the design of clinical trials that enable or facilitate extrapolation.

[29-P6-02] A Comparison of Meta-Analytic Methods for Synthesizing the AUC and Other Probabilistic Superiority Measures

Presenter: Bethany Hamilton Bhat

Authors: Bethany Hamilton Bhat [1]; Wolfgang Viechtbauer [1]

Affiliations: [1] Maastricht University, Maastricht

Introduction

Measures of probabilistic superiority (PS), including the area under the receiver operating characteristic curve (AUC), the common language effect size (CLES), and the concordance probability, express the probability that a randomly chosen subject from one group has a higher outcome value than a randomly chosen subject from another group. These measures are related to the Mann-Whitney U statistic and have attractive properties for applied researchers, including intuitive interpretation and robustness to unequal variances. However, PS measures are bounded between 0 and 1 and have skewed sampling distributions, both of which pose challenges for meta-analytic pooling. Methods for meta-analytically synthesizing PS measures have not been properly examined, and we focus on the AUC as a representative PS measure. It remains unclear whether raw PS measures should be pooled directly or whether a transformation is necessary, and which sampling variance formula performs best as inverse-variance weights in random-effects meta-analysis of PS.

Methods

Using individual patient data from five randomized controlled trials comparing Uro-Vaxom to placebo for urinary tract infection recurrence (Huber et al., 2000), we computed the nonparametric and parametric AUC estimates under homoskedastic and heteroskedastic assumptions. We evaluated 15 sampling variance formulas spanning distribution-free and normal-based approaches. Pooling methods included direct pooling of raw AUC estimates and pooling of transformed estimates using logit, probit, and arcsine transformations with regular and integral back-transformations.

Results

Pooled AUC estimates were highly consistent across sampling variance formulas and similar across pooling approaches. Direct pooling yielded estimates of 0.669 to 0.685 logit: 0.663 to 0.680 probit: 0.665 to 0.681 and arcsine: 0.666 to 0.683, indicating minimal differences among methods.

Discussion

The meta-analytic pooling method had little impact on the pooled estimates in this dataset, with all approaches yielding similar results. This suggests that meta-analytic conclusions based on AUC may be robust to the choice of transformation, AUC specification, and sampling variance formula.

Time permitting, the results of a Monte Carlo simulation study will be presented to examine whether this pattern holds more generally. Parallels with the meta-analysis of standardized mean differences will be discussed, along with practical recommendations for applied researchers.

[29-P6-03] Evaluation of Meta-Analysis Methodologies for Small-Scale and High-Heterogeneity Data: A Meta Analysis of Gender-Affirming Hormone Therapy

Presenter: Yota Funakoshi

Introduction

Gender-affirming hormone therapy (GAHT) plays a crucial role in improving the quality of life for transgender individuals however, its long-term effects on cardiovascular disease, thrombosis, and bone metabolism remain unclear. This uncertainty arises from limited research participation due to social stigma and restricted healthcare access, resulting in small-scale, single-center studies and the risk of conflating 'absence of evidence' with 'evidence of absence' regarding clinical risks. Meta-analysis provides a valuable approach to detect subtle risks and generate quantitative evidence by integrating multiple studies. However, heterogeneity due to differences in hormone regimens, administration methods, confounding factors (e.g., smoking, stress), publication bias, and short follow-up periods necessitates careful interpretation.

Methods

To address limited data and high inter-study heterogeneity, this study employs multiple statistical approaches with sensitivity analyses. Although random-effects models were prioritized to account for clinical and methodological heterogeneity, results were compared with fixed-effect models to assess the robustness of estimates across different weighting schemes. Inter-study variance (τ^2) was estimated using both the DerSimonian-Laird (DL) and Restricted Maximum Likelihood (REML) methods. Although DL is widely used, it may underestimate τ^2 under high heterogeneity, whereas REML provides more reliable estimates. To ensure conservative and reliable inference, we employed the Hartung-Knapp-Sidik-Jonkman (HKSJ) adjustment and Kenward-Roger (KR) correction, which provide superior coverage probabilities in the presence of small study numbers and high heterogeneity. Publication bias was assessed using funnel plots, Egger's test, and selection models. Additionally, prediction intervals were calculated to estimate the range of effects expected in future studies.

Results

The meta-analysis evaluated the impact of GAHT on lipid profiles. Focusing on LDL cholesterol in female-to-male (FTM) individuals, a significant increase was observed: +6.50 mg/dL (95% CI: 3.70-9.31, $p<0.001$) under the fixed-effects model and +8.07 mg/dL (95% CI: 1.58-14.56, $p=0.015$) under the random-effects model. Despite the significant mean increase, substantial heterogeneity was observed ($I2=77.9$), suggesting that the effect size is highly context-dependent across different clinical settings and regimens. While GAHT appears to be associated with an elevation in LDL cholesterol, the wide prediction intervals and observed heterogeneity necessitate a nuanced, individualized approach to cardiovascular risk management in this population.

Discussion

Discussion will be presented at the conference.

[29-P6-04] The Impact of the Choice of the Likelihood Function in Bivariate Meta-Analysis of Time to Event Data: An Application to Surrogate Endpoint Evaluation

Presenter: Ireoluwa Adeleke, Janharpreet Singh, Sam Khan, Richard Riley, Sylwia Bujkiewicz

Authors: Ireoluwa Adeleke [1]; Janharpreet Singh [1]; Sam Khan [1]; Richard Riley [2]; Sylwia Bujkiewicz [3]

Affiliations: [1] University of Leicester, Leicester; [2] University of Birmingham, Birmingham; [3] Biostatistics Research Group, Division of Public Health & Epidemiology, School of Medical Sciences, University of Leicester, Leicester

Introduction

Surrogate endpoints are used in clinical trials to provide early evidence of treatment effects on final outcomes. Bivariate meta-analysis allows joint analysis of treatment effects on a surrogate endpoint and a final outcome across studies. The standard bivariate meta-analytic model uses aggregate study-level effectiveness estimates, assuming a bivariate normal likelihood at within- and between-study levels. However, in rare diseases where data are sparse, the assumption that treatment effects on log scales (log hazard ratio (HR), log odds ratio (OR), or log incidence rate ratio (IRR)) are normally distributed may not hold, which could affect assessment of surrogate endpoint validity.

Methods

We used the standard Daniels and Hughes model, assuming the effects on log scales follow a normal likelihood. We also considered alternative likelihood specifications for proportions and counts of events using independent Binomial (D&H-IB) and Poisson (D&H-IP) models respectively. Impact of ignoring the within-study correlation was explored using D&H model with independent normal likelihoods (D&H-IN). To explore the effect of the choice of likelihood function on the surrogacy parameter estimates, we applied the models to a case study in cancers with mismatch repair deficient (dMMR) biomarker status (colorectal, endometrial and gastric) and their proficient (pMMR) counterparts, separately and together, with progression-free survival as a surrogate endpoint for overall survival.

Results

Across the D&H, D&H-IN and D&H-IP models, greater uncertainty was observed in the dMMR data. For example, the slope estimate for the D&H-IN model in dMMR was 0.42 (95% credible interval: -0.78, 1.55), compared with 0.72 (0.36, 1.16) in pMMR and 0.76 (0.49, 1.08) in the combined dataset. In contrast, the D&H-IB model produced smaller slope estimates, for example 0.15 (-0.63, 1.18) in dMMR, compared with 0.42 (-0.78, 1.55) under D&H-IN. This pattern was consistent across all datasets, with greater uncertainty particularly in dMMR, representing the rare cancer subtype.

Discussion

The choice of within-study likelihood in Bayesian bivariate meta-analysis influences on surrogate endpoint evaluation, especially in rare cancers where data are limited. In more common cancers with larger sample sizes, results are generally more consistent across different models. The performance of these models will be further assessed using through cross-validation and simulation studies.

[29-P6-05] Bayesian synthesis of timing-risk associations: an individual patient data meta-analysis to identify the optimal timing of anticoagulation after ischaemic stroke

Presenter: Konstantina Chalkou, Georgia Salanti, o n behalf of the Catalyst collaborators

Authors: Konstantina Chalkou [1]; Georgia Salanti [1]

Affiliations: [1] Institute of Social and Preventive Medicine, University of Bern, Bern

Introduction

In patients with acute ischaemic stroke and atrial fibrillation, initiating direct oral anticoagulants (DOACs) within 4 days seems to reduce the risk of recurrent ischaemic stroke, symptomatic intracerebral haemorrhage, or unclassified stroke within 30 days. However, the optimal timing for DOAC initiation within this early window remains unclear Methods

We conducted an individual patient data meta-analysis of four randomized clinical trials comprising 6 619 participants. We developed a hierarchical synthesis model to estimate the association between DOAC initiation timing (days from symptom onset for index ischemic stroke) and the risk of a composite outcome (recurrent ischaemic stroke, symptomatic intracranial haemorrhage, or unclassified stroke) within 30 days.

Methods

We implemented four synthesis models: two for the targeted randomised timing ("as randomised") and two for the observed timing ("as observed"). In each approach, we used a linear and a non-linear (spline-based) function to estimate the association between timing and outcome risk. Model parameters were synthesized across studies using random-effects meta-analysis fitted within a Bayesian framework.

Results

Across all models, we found that delaying DOAC initiation was consistently associated with a higher 30-day risk of the composite outcome. The time-risk analyses showed that each additional day of delay increased the probability of the composite outcome. The non-linear models suggested an almost linear relationship, indicating that even starting DOACs on day 1 did not appear to increase harm and might offer the greatest benefit. This continuous, dose-response approach offers a more informative perspective than traditional analyses based on arbitrary time categories (for example, ≤4 days vs >4 days), which can hide the underlying time-risk relationship.

Discussion

This IPD-based Bayesian synthesis enables modelling the association between treatment timing and risk, overcoming the limitations of arbitrary cutoffs in traditional meta-analyses. The approach integrates data across multiple trials and quantifies uncertainty in both linear and non-linear risk relationships. Beyond the clinical application to anticoagulation timing, the methodology provides a generalizable framework for evaluating temporal treatment effects in clinical research using individual-level data.

[29-P6-06] Effectiveness and safety of opioid substitution and antagonist therapies for opioid use disorder: A systematic review and network meta-analysis

Presenter: Myanca Rodrigues

Authors: Myanca Rodrigues [1]; Brittany B. Dennis [2]; Behnam Sadeghirad [3]; Anna Oprea [4]; Mikail Malik [5]; Leen N Naji [6]; Alannah McEvoy [4]; Devon Malhotra [4]; Sameer Parpia [3]; Lehana Thabane [3]; Zainab Samaan [1]

Affiliations: [1] Queen's University, Kingston; [2] University of British Columbia, Vancouver; [3] Departments of Medicine and Health Research Methods, Evidence, and Impact, McMaster University, Hamilton; [4] McMaster University, Hamilton; [5] University of Toronto, Toronto; [6] University of Arizona, Phoenix, AZ

Introduction

Opioid use disorder (OUD) is a public health crisis in North America. Comparative effectiveness evidence for opioid substitution and antagonist therapies (OSATs) is needed to inform clinical and policy decisions. Our study aim was to assess the relative benefits and harms of OSATs for OUD through a network meta-analysis (NMA) of randomized controlled trials (RCTs).

Methods

We searched Medline, EMBASE, PubMed, PsycINFO, Web of Science, Cochrane Library, Cochrane Clinical Trials Registry, and the National Institutes of Health Clinical Trials Registry through August 10, 2025. Trials assessing the effectiveness and safety of any OSAT for people with OUD were eligible for inclusion. Pairs of reviewers independently screened studies, extracted data, and assessed risk of bias. A frequentist random-effects model was used for all NMAs, and certainty of evidence for treatment effects was assessed using the Grading of Recommendations Assessment, Development, and Evaluation approach. Our outcomes included: treatment retention (people remaining in follow-up at study end), opioid abstinence (during follow-up), criminal activity, all-cause mortality, and overdose-related death.

Results

We included 103 RCTs (n=26,624 people with OUD) assessing 13 interventions. Compared with placebo, moderate-certainty evidence demonstrated that heroin (Odds Ratio [OR]=31.47 95% confidence interval [CI]=12.29,80.58), was most effective for treatment retention. Heroin + methadone (OR=8.00 95%CI=3.21,19.93) and high-dose methadone (OR=6.94 95%CI=3.76,12.82) also ranked highly. High- and low-dose buprenorphine (moderate certainty) and high-dose methadone (moderate certainty for retention, low certainty for abstinence), improved abstinence as measured by urine drug screens. No intervention reduced criminal activity (moderate certainty), all-cause mortality (low certainty), or overdose-related death (low certainty).

Discussion

Heroin-assisted therapy, alone or with methadone, and high-dose methadone were the most effective interventions for treatment retention, supported by moderate-certainty evidence. Buprenorphine and high-dose methadone improved abstinence, although certainty was lower. Evidence for mortality, overdose, and other patient-important outcomes remains limited. Heroin-assisted therapy exclusively enrolled people with heroin dependence, and findings should not be taken for its initiation in those without prior use. Larger trials directly comparing active interventions, particularly methadone, buprenorphine, and heroin-assisted therapy, and enrolling patients with fentanyl use, are needed to guide treatment in the context of an evolving toxic drug supply.

[29-P6-07] Indirect treatment comparisons in 15 years of AMNOG

Presenter: Janik Beuermann

Authors: Janik Beuermann [1]; Vera Grafe [1]; Mustafa Rmo [1]

Affiliations: [1] Ecker + Ecker GmbH, Hamburg

Introduction

Under the German Act on the Reform of the Market for Medicinal Products (AMNOG), which came into force in 2011, new medicines must undergo an early benefit assessment upon entering the market. The aim is to determine whether there is any evidence for an additional benefit compared to existing standard therapies. Where possible, this comparison should be based on direct randomised controlled trials (RCTs). If such data are unavailable, pharmaceutical companies may include indirect evidence instead. However, as evidence from indirect treatment comparisons (ITC) is less robust than data from RCTs, gaining acceptance for it in the early benefit assessment by the Federal Joint Committee (G-BA) is challenging. Here we ask, what trends regarding indirect comparisons can be identified from 15 years of AMNOG?

Methods

All ITCs are identified by screening the G-BA's resolution justifications since 2011 (excluding orphan drug assessments) and are classified according to recognition, methodology, and therapeutic area. The study examines whether a temporal trend can be identified in the proportion of procedures involving ITCs, and within this, the proportion of recognised ITCs. Furthermore, we identify procedures that can be regarded as important milestones for ITCs in early benefit assessment, e.g. first-time recognition or derivation of substantial additional benefit.

Results

During the 15-year period of the AMNOG, approximately 850 non-orphan assessments were conducted, 162 of which included indirect evidence. Of these, the G-BA recognised additional benefit in only 15 assessments, and accepted indirect evidence in 44. Key milestones include procedure 143, in which the G-BA awarded considerable additional benefit based on an unadjusted comparison, procedure 386, in which an ITC according to Bucher served as the sole evidence for an additional

benefit for the first time, and procedure 1190, in which an ITC calculated according to Bucher based purely on published data led to considerable added benefit.

Discussion

Although they are rarely leading to an additional benefit granted by G-BA, ITCs are an important strategic tool for pharmaceutical manufacturers. It will be interesting to observe the influence that the EU HTA launched in 2025 will have on ITCs within the AMNOG in the coming years.

[29-P6-08] Reducing animal use in long-term carcinogenicity studies by using historical control data

Presenter: Max Menssen

Authors: Max Menssen [1]; Carsten Kneuer [2]; Gyamfi Akyianu [2]; Christian Röver [1]; Tim Friede [1]; Frank Schaarschmidt [3]

Affiliations: [1] Universitätsmedizin Göttingen, Institut für Medizinische Statistik, Göttingen; [2] Bundesinstitut für Risikobewertung, Fachgruppe Toxikologie der Wirkstoffe und ihrer Metabolite, Berlin; [3] Leibniz Universität Hannover, Institut für Zellbiologie und Biophysik, Abteilung Biostatistik, Hannover

Introduction

For ethical and economic reasons, the reduction of animal use in pre- and non-clinical toxicology is highly desirable. We will demonstrate possible sample-size reductions of up to 80% in control groups of long-term carcinogenicity studies (OECD test guidelines 451 and 453) by using historical control data.

Such studies typically include one control group and at least three treatment groups (50 animals each) for each sex, to which the compound of interest is applied in increasing dosage. At the end of the study, the numbers of observed histopathological findings (e.g. tumors) are compared between the treatment groups and the control in a Dunnet-type test procedure.

Methods

In order to augment animals in the current control group with historical control data, we propose two approaches that rely on dynamic Bayesian borrowing (via robustified prior distributions) and rank-based simultaneous credible intervals for risk ratios.

Results

Monte-Carlo simulations revealed that the proposed approaches enable a substantial reduction of the control group's sample size, while still controlling the FWER up to a satisfactory level and retaining the power of the test. Furthermore, it can be shown that robustification of priors enables a partial protection against possible drift (heterogeneity) between current and historical observations.

Discussion

This highlights the potential of Bayesian study designs to reduce animal use in toxicology through re-use of the large pool of existing control data. Even though the methodological development was driven by toxicological applications, the proposed methodology might also be of interest in the context of clinical multi-arm studies.

[29-P6-09] Evaluating Ensemble Feature Selection Methods: A Systematic Review and Insights on Stability and Reliability

Presenter: Robin Martens

Authors: Robin Martens [1]; Gökhan Gül [1, 2, 3]; Vincent ten Cate [1, 2, 3]; Lisa Hartung [4]; Philipp S. Wild [1, 2, 3, 5]

Affiliations: [1] Preventive Cardiology and Preventive Medicine, Department of Cardiology, University Medical Center of the Johannes Gutenberg University Mainz, Mainz; [2] German Center for Cardiovascular Research (DZHK), partner site Rhine Main, Mainz; [3] Clinical Epidemiology and Systems Medicine, Center for Thrombosis and Hemostasis, University Medical Center of the Johannes Gutenberg University Mainz, Mainz; [4] Institute of Mathematics, Johannes Gutenberg University Mainz, Mainz; [5] Institute of Molecular Biology (IMB), Mainz

Introduction

Advances in biochemical technologies have produced high-dimensional datasets with many predictors, posing challenges for statistical analysis. In this setting, variable selection aims to identify a subset of predictors that capture effects on specific outcomes. However, individual methods often yield unstable feature subsets, due to sensitivity to sample variation, measurement noise, and model assumptions. Ensemble feature selection has been proposed to address this limitation by aggregating results across methods.

Methods

A systematic review was conducted using broad search terms to account for inconsistent terminology in ensemble feature selection, yielding ~86,000 records. Machine-learning-assisted abstract screening was performed using ASReview, documenting all screening decisions to ensure transparency and reproducibility. Based on identified methods, a simulation study evaluated stability (using the Nogueira stability measure), false discovery rate (FDR), and true positive rate (TPR) across different ensemble structures compared with their corresponding base methods. To ensure feasibility, the analysis focused on the five most cited filter, wrapper, and embedded methods combined with the five most common aggregation strategies, including score-, rank-, and decision-based approaches.

Results

The review identified 262 relevant papers on ensemble feature selection, comprising 184 distinct feature selection methods, three ensemble architectures, and 86 aggregation strategies.

Simulation results showed that ensemble feature selection improved stability primarily for base methods with low initial stability (<0.3). In some cases, stability more than doubled, for example, for the top 50 ranked features from SVM-RFE, stability increased from 0.13 to 0.40 using ensemble methods. Although ensemble variants often outperformed base methods in either FDR or TPR, individual methods generally showed more balanced performance. For instance, maximum aggregation in homogeneous ensembles achieved higher TPR (0.81-1.00 vs. 0.32-0.98) but also higher FDR (0.88-0.92 vs. 0.54-0.91).

Discussion

Although base feature selection methods often provide balanced trade-off between TPR and FDR, ensemble approaches allow variable selection to be tuned toward specific objectives. Additionally, ensemble methods may serve as a useful post-evaluation step to confirm reliable predictors or as a preselection step to reduce dimensionality while retaining relevant variables, thereby improving the robustness of subsequent analyses.

[29-P6-10] Machine Learning Models for Predicting Adverse Birth Outcomes: A Systematic Review and Risk of Bias Assessment

Presenter: Daisy Chelangat

Authors: Daisy Chelangat [1, 2]; Faith Neema [1]; Akbar Waljee [3]; Amina Abubakar [1]; Inigo Bermejo Delgado [2]

Affiliations: [1] Aga Khan University, Nairobi; [2] Hasselt University, Hasselt; [3] University of Michigan, Michigan

Introduction

Early identification of women and infants at risk of adverse birth outcomes is essential to inform timely interventions, improve maternal and neonatal care, and reduce long-term morbidity. Many prediction models developed for this purpose are at a high risk of bias, limiting their reliability and applicability in clinical practice.

Objectives

We conducted a systematic review of Machine Learning (ML)- based predictive models for adverse birth outcomes (low birth weight, preterm birth, and stillbirth) to critically assess their methodological rigor and identify top predictors.

Methods

Peer-reviewed journal articles published between January 2014, and November 2024 were systematically retrieved from three databases: Web of Science, PubMed, and Scopus. The quality, applicability, and risk of bias in the studies were accessed using the Prediction model Risk Of Bias Assessment Tool for Artificial Intelligence (PROBAST+AI).

Results

A total of 88 articles were included. Overall, 54% (46/86) of the included studies were rated as having a high concern regarding the quality of model development, and 72% (61/85) had a high risk of bias in model evaluation. Methodological issues identified were insufficient sample size (56%), use of univariate feature selection methods without regularization, lack of calibration after handling data imbalance, and poor handling of missing data. The top predictors for poor birth outcomes across studies were maternal anthropometric measures, age, educational level, gestational diabetes, hypertension and number of children. Logistic regression remains the most popular technique.

Discussion

Most ML-based prediction modelling studies for adverse birth outcomes are at a high risk of bias and have high concerns about the quality of model development.

Systematic review registration

This protocol for this review was registered with the PROSPERO register of systematic reviews (Registration ID: CRD42024611652).

[29-P6-11] A critical review of clinical prediction models in oncology: is this SEERious research?

Presenter: Biruk Tsegaye

Authors: Biruk Tsegaye [1]; Jie Ma [1]; Jaga Fryc [1]; Zahina Yasmin [1]; Melissa Brookfield [1]; Richard Riley [2, 3]; Paula Dhiman [1]; Gary Collins [2, 3]

Affiliations: [1] Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology & Musculoskeletal Sciences, Botnar Research centre, University of Oxford, oxford, United Kingdom, Oxford; [2] Department of Health Sciences, School of Health Sciences, college of Medicine and Health, University of Birmingham, Birmingham, United Kingdom, Birmingham; [3] NIHR Birmingham Biomedical Research Centre, Birmingham, University of Hospital Birmingham NHS Foundation Trust and University of Birmingham, Birmingham, United Kingdom, Birmingham

Introduction

Clinical prediction models estimate an individual patient's probability of current or further health outcomes conditional on their values of multiple predictors, including demographic and clinical characteristics, laboratory results, imaging findings, and other diagnostic data.

The surveillance, epidemiology, and End Results (SEER) cancer registry collects and publishes comprehensive patient-level data on cancer incidence, treatment, and survival across the United States population. most models developed using the SEER database have not been implemented in routine clinical practice. We aimed to characterise and examine the scale of clinical prediction models developed using the SEER cancer registry reflects formulaic research practices whilst having limited scientific value.

Methods

We conducted a systematic review of clinical prediction model development studies using SEER. We searched MEDLINE via PubMed for studies published between January 1st and August 10th, 2023. Studies were included if they developed a clinical prediction model using two or more predictors. Data extraction included study context, intended use, reporting guideline adherence, external validation practices, and the rationale provided for using SEER.

Results

Of 436 included studies, 94.5% had China-affiliated first authors, while only 3.2% were from the United States. Most models focused on prognosis (87%) studies, Cox regression was the predominate modelling approach. Most models were presented as a nomogram (77%). Adherence to reporting guidelines was uncommon (12.3%), and few studies mentioned a study protocol (3%). Only 6% of China-affiliated studies reported the models intended clinical settings, and only 3% specified the intended countries of use. Few studies mentioned the suitability of using SEER (13%). Among China-affiliated studies, 22% used non-US data for external validation and 18% recommended non-US data for such validation.

Discussion

Most of the included clinical prediction models developed using SEER lack a clearly stated rationale, defined target population, intended clinical settings, and failed to address whether the SEER data were appropriate for non-US populations. Many studies demonstrate characteristics of formulaic research and producing output of limited scientific value. Researchers should rigorously justify their data sources, comply with the TRIPOD+AI reporting guideline, avoid concluding that the SEER-based models are applicable outside of the US without conducting appropriate external validation.

[29-P6-12] Prediction model developed on the basis of Aggregate data Meta-analysis: a systematic review and methodological summaries

Presenter: Junfeng Wang

Authors: Junfeng Wang [1]; Shengfeng Wang [2]

Affiliations: [1] Maastricht University, Maastricht; [2] Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing, China., 中国 北京

Introduction

Prediction models derived from Aggregate data Meta-analyses of published aggregate data were increasingly applied in health research. However, methodological limitations and incomplete reporting posed significant challenges to their development and clinical application.

Methods

A systematic search of PubMed, Embase, Cochrane Library, and Web of Science was conducted up to September 2025 to identify studies developing prediction models based on aggregate data Meta-analyses. Key data were extracted on study characteristics, handling of original studies, model development, validation, and clinical utility. Findings were synthesized using a narrative approach.

Results

40 studies met inclusion criteria. Of these, 35 (87.5%) performed both model development and validation, 24 (60.0%) reported Meta-analysis registration, and 37 (92.5%) assessed original study quality. Concerning original studies and effect estimates, none (100%) distinguished between derivation and validation studies, 36 (90.0%) did not address multiple studies from the same data source, 7 prognostic studies (92.5%) did not report handling of inconsistent outcome time points, and 29 (72.5%) did not differentiate univariable from multivariable effect estimates.In model development, 29 studies (72.5%) handled predictor selection inappropriately 19 (47.5%) did not report harmonization of different effect measures 20 (50.0%) did not address different expressions of the same effect measure 16 (40.0%) did not handle continuous variables 39 (97.5%) did not assess linear or nonlinear relationships between predictors and outcomes 34 (85.0%) did not address predictor collinearity and 36 (90.0%) did not consider interaction effects.Regarding validation and clinical utility, 15 studies (37.5%) evaluated calibration without reporting intercepts, 13 (32.5%) conducted decision curve analysis without intercept information, and 3 (7.5%) performed decision curve analysis without prior calibration assessment. Stratification by presence of statistical experts indicated a higher proportion of appropriately addressed methodological issues in teams with expertise.

Discussion

Aggregate data Meta-analytic methods for prediction model development remained underdeveloped, especially in differentiating and integrating original studies. Fundamental issues in model development were often overlooked, and the pathway from Meta-analysis to prediction model remained unclear. Conducting validation and clinical utility assessments without complete model reporting may have compromised reliability and overestimated clinical applicability. Engagement of methodological experts substantially improved study quality.

[29-P6-13] The R package gam2formula, or: How do I get the equation of an estimated smooth?

Presenter: Stefan Gehrig

Authors: Paul Bach [1]; Stefan Gehrig [1]

Affiliations: [1] Medical Biometry and Statistics Unit, Institute for Quality Assurance and Transparency in Healthcare (IQTIG), Berlin, Germany, Berlin

Introduction

The popular R package mgcv allows the estimation of smooth effects for continuous predictor variables in a generalized additive mixed model (GAMM) framework, providing a variety of smoothers

with penalized likelihood maximization. Yet, there is currently no easy-to-use functionality in the R universe to derive the analytical expression for the estimated smooth conditional mean function. This limits transparency and reproducibility in many clinical applications. For example, model predictions for new patients can currently only be computed using the original fitted model object. At the Institute for Quality Assurance and Transparency in Healthcare (IQTIG), this presented a barrier for using GAMMs in risk adjustment models for healthcare quality assurance.

Methods

We built an open source R package that delivers closed formulaic representations for many common smoothers from mgcv and can be applied in a fast and simple manner. The derivation is exact and relies on a change of basis of the internally used empirically centered basis functions.

Results

The R package gam2formula prints formulas for many of the most commonly used smoothers, converts them into LaTeX regression tables, and makes predictions on new data without access to the training data or original model object.

Discussion

Our tool circumvents typical barriers for the application of advanced smoothing methods in clinical applications that require a high degree of transparency (e.g., the detailed reporting of regression coefficients) and reproducibility (e.g., the computation of predictions independent of access to confidential training data the transfer of prediction models between software platforms). By revealing the expressions behind the prediction machinery, it can also support educators who teach semi-parametric modeling of smooth effects. We discuss and invite possible extensions to the software.

[29-P6-14] MetaNB: An R Package for Meta-Analysis of Net Benefit and Value-of-Information

Presenter: Kim Zhipei Wang

Authors: Kim Zhipei Wang [1]; Junfeng Wang [1]; Laure Wynants [1, 2]

Affiliations: [1] Maastricht University, Maastricht; [2] KU Leuven, Leuven

Introduction

Clinical prediction models should be validated across multiple settings before their clinical usefulness can be established. Evidence-synthesis methods for external validation studies are well known for predictive performance such as discrimination and calibration, but not yet for clinical utility. Net benefit provides a decision-analytic measure of clinical utility. When only study-level performance measures are available, net benefit across settings can be synthesized using a trivariate random-effects meta-analysis of prevalence, sensitivity, and specificity. However, accessible software for implementing these methods and for quantifying decision uncertainty through value-of-information analysis is currently lacking.

Methods

We developed MetaNB, an R package that performs Bayesian trivariate random-effects meta-analysis using Markov chain Monte Carlo sampling via JAGS. The package accepts study-level data and produces posterior estimates of clinical utility, including pooled net benefit and its credible interval, prediction interval for new settings, and the posterior probability that a model is clinically useful in a new setting, for given relevant harm-to-benefit ratios. Building on these quantities, the package further estimates metrics of uncertainty for the decision to implement the model in care, including expected value of perfect and partial perfect information in heterogeneous settings.

Results

The implementation is illustrated using a case study on meta-analytic findings from the validation of a prediction model to diagnose ovarian cancer, and provides a framework for estimating net benefit across heterogeneous settings and associated value-of-information measures. Functionality is illustrated through vignettes demonstrating estimation, visualization, and interpretation of clinical utility and value-of-information measures in multicenter validation contexts. These examples show the types of outputs the package generates, and provide guidance on how to interpret these results.

Discussion

By translating previously proposed methodology into an accessible R implementation, MetaNB enables practical evidence synthesis of clinical utility and principled quantification of decision uncertainty. The package supports informed decisions about model adoption and the potential value of further validation studies, thereby promoting transparent evaluation of clinical prediction models across settings.

[29-P6-15] The potential for improvement of health care providers: How many adverse events can be prevented?

Presenter: Johannes Rauh

Authors: Johannes Rauh [1]

Affiliations: [1] Institut für Qualitätssicherung und Transparenz im Gesundheitswesen, Berlin

Introduction

Within the eQS programme, managed by the IQTIG, the quality of care of German health care providers is measured and monitored using almost 300 quality indicators. Each indicator operationalizes a quality requirement on a process or on treatment outcomes. The indicator results of individual providers are used both to initiate quality improvement measures and for public reporting. In addition, the totality of the indicator results of all providers together allows to describe quality of health care from a system level point of view. One quantity of interest is the potential for improvement: How many adverse events could possibly be prevented by suitable quality improvement measures? This quantity is related to the question what level of quality is achievable.

Methods

An early approach to estimate the achievable quality of care is called ABC ("Achievable Benchmarks of Care", [2]). Since the ABC approach does not take into account patient specific confounders (i.e. risk adjustment), it is not suitable for many quality indicators, in particular outcome indicators. Moreover, the ABC approach combines two statistical models: A Bayesian model is used to identify benchmark providers. The level of quality of these benchmark providers is then estimated using a frequentist model. The authors argue that this combination of two models improves face validity and makes it easier to communicate results. On the other hand, statistical inference and quantification of uncertainty are difficult when combining different models.

Results

Therefore, the IQTIG developed a new methodology based on a single mixed model, in which treatment quality is modeled as a random effect.

Discussion

In this talk, I will compare both approaches and explain some modeling choices. I will also summarize the experiences from applying our approach.

References:

[1] IQTIG, 2024: Methodische Grundlagen, Version 2.1, Anhang B. https://iqtig.org/das-iqtig/wie-wir-arbeiten/grundlagen/methodische-grundlagen/

[2] Weissman, NW Allison, JJ Kiefe, CI Farmer, RM Weaver, MT Williams, OD et al. (1999): Achievable benchmarks of care: the ABC(TM)s of benchmarking. Journal of Evaluation in Clinical Practice 5(3): 269-281. DOI: 10.1046/j.1365-2753.1999.00203.x.

[29-P6-16] A systematic review of randomized controlled trials on the use of artificial intelligence in primary care: Preliminary results

Presenter: Lukas M. Horstmeier

Authors: Urs Alexander Fichtner [1]; Angelika Eisele-Metzger [2]; Carola Fichter [3]; Anna Kabalimu [3]; Vanessa Kaiser [4]; Claudia Salm [5]; Sebastian Voigt-Radloff [6]; Lukas M. Horstmeier [5]; Erik Farin-Glattacker [7]; Andy Maun [8]

Affiliations: [1] Medical Faculty and Medical Center - University of Freiburg, Freiburg; [2] Uniklinik Freiburg, Institut für Evidenz in der Medizin (IfEM), Freiburg im Breisgau; [3] Uniklinik Freiburg, Institut für Allgemeinmedizin (IfA), Freiburg im Breisgau; [4] Universitätsklinikum Freiburg, Freiburg im Breisgau; [5] Universitätsklinikum Freiburg, Freiburg; [6] Uniklinik Freiburg, Institut für Biometrie und Statistik (IMBI), Sektion Versorungs- und Rehabilitationsforschung, Freiburg im Breisgau; [7] Section of Health Care Research and Rehabilitation Research, Institute of Medical Biometry and Statistics, Medical Faculty and Medical Center - University of Freiburg, Freiburg im Breisgau; [8] Institute of General Practice/Family Medicin, Medical Faculty and Medical Center - University of Freiburg, Freiburg im Breisgau

Introduction

Clinical applications using Artificial Intelligence (AI) is rising rapidly, however, robust evidence on patient-related effects is rather limited. Specifically, to our knowledge, the application in primary care.

The aim of this systematic review is to identify effects of AI-assisted interventions on patient-relevant outcomes and care processes. in primary care in recently published randomized controlled trials (RCTs).

Methods

Included were RCTs published from 2021 onwards, investigating AI-supported tools used in primary care for patients with defined health problems or their practitioners. Excluded were studies that did not establish a direct link to patient-relevant outcomes (e.g., diagnostic optimization by AI alone). The literature research was conducted without language restrictions in CENTRAL, CINAHL, Embase, and MEDLINE. The two-step screening process, data extraction, and RoB-2 risk of bias assessment are executed by at least two independent researcher respectively. A narrative synthesis is planned due to expected heterogeneity. The literature search was conducted in March 2025 and yielded 2019 enties. During title/abstract screening, 1,269 studies were excluded full-text screening is currently running.

Results

Final results will be presented at the congress.

Discussion

Our review focuses of primary care sheds light on a currently less characterized area of AI application. The focus on primary care relevant populations and outcomes will provided a differentiated practical assessment on the utility of AI.

This systematic review expands evidence on AI-usage in clinical practice by summarizing patient-relevant effects of AI-assisted interventions in primary care. The results aim to provide a basis for the evaluation, implementation, and further development of AI applications in the general practice setting.

P7 - AI and Biomedical Data

Poster area: P7

[29-P7-01] Olink Proteomics Data: A Comparison of Statistical Strategies and Future Challenges

Presenter: Arianna Galotta

Authors: Arianna Galotta [1]; Marta Brambilla [2]; Francesco Maria Mattio [2]; Cristina Banfi [2]; Marina Camera [2, 3]

Affiliations: [1] IRCCS Centro Cardiologico Monzino, Milano; [2] IRCCS Centro Cardiologico Monzino, Milan; [3] Università degli Studi di Milano, Milan

Introduction

Olink® proteomics platforms enable high-throughput biomarker discovery through multiplex protein quantification. In cardiovascular research, these technologies offer opportunities to identify predictive biomarkers. However, the resulting datasets are typically high-dimensional, with many proteins measured relative to the sample size, requiring appropriate statistical approaches. This study aimed to evaluate and compare statistical methods for analysing Olink proteomics data from patients with coronary artery disease, with the goal of identifying plasma biomarkers associated with cardiovascular mortality.

Methods

A total of 69 plasma samples were analysed, including 17 patients (24.6%) who experienced cardiovascular mortality. Protein expression levels were measured using four Olink Target 96 panels (cardiometabolic, cardiovascular II and III, and inflammation), yielding 333 Normalized Protein eXpression (NPX) variables. The analytical workflow included univariate t-tests, principal component analysis (PCA), Gene Set Enrichment Analysis (GSEA), heatmap visualization, Boruta-based feature selection, and multivariate logistic regression with stepwise selection. Analyses were conducted using SAS v9.4 and R v4.3.1, incorporating the OlinkAnalyze R package [Nevola K. et al. OlinkAnalyze: Facilitate Analysis of Proteomic Data from Olink. R package version 4.2.0, 2025].

Results

Univariate analyses did not identify statistically significant differences after multiple testing correction, as confirmed by volcano plots. PCA showed limited explanatory power of the first two components, indicating poor separation between outcome groups. Similarly, GSEA and heatmap analyses did not reveal significant enrichment patterns. In contrast, the Boruta algorithm identified relevant features, which were included in a multivariate logistic regression model. Stepwise selection based on unadjusted p-values produced a predictive model with good discriminative performance (AUC = 0.89, 95% CI: 0.81-0.97). Clinical expertise was essential for interpretation.

Discussion

These findings highlight the importance of combining multiple statistical approaches in high-dimensional settings. While no single method provided definitive results, their integration enabled identification of potentially relevant biomarkers and development of a predictive model. The small sample size remains a key limitation, affecting stability and reproducibility. Future studies should explore synthetic data generation to emulate larger datasets and improve robustness in small-scale Omics research.

[29-P7-02] Screening Non-Monotonic Expression Profiles in RNA-Seq via Complementary ROC Indices

Presenter: Dincer Goksuluk

Authors: Dincer Goksuluk [1]; Merve Basol Goksuluk [1]; Ebru Ozturk [2]; Unal Erkorkmaz [1]; Asuman Deveci Ozkan [3]; Hamdi Furkan Kepenek [2]

Affiliations: [1] Department of Biostatistics, Faculty of Medicine, Sakarya University, Sakarya; [2] Department of Biostatistics, Faculty of Medicine, Hacettepe University, Ankara; [3] Department of Medical Biology, Faculty of Medicine, Sakarya University, Sakarya

Introduction

RNA-Seq differential expression analysis is widely used for biomarker discovery and disease characterization. However, standard methods mainly target monotonic mean shifts and may miss clinically relevant molecular profiles in which both low and high expression levels are associated with disease. Such non-monotonic, "improper" profiles may reflect biological heterogeneity and limit detection of candidate biomarkers.

Methods

We evaluated two ROC-derived indices, generalized area under the ROC curve (gAUC) [1] and length of the ROC curve (LROC) [2], as complementary screening tools to the widely used differential expression method DESeq2 [3]. In 18 simulation scenarios based on negative binomial count models, varying sample size and overdispersion, methods were compared using ranking-based true positive rate and positive predictive value, with emphasis on improper profiles. We then applied the approach to a public cervical cancer miRNA sequencing dataset including 29 tumor and 29 non-tumor samples. Candidate improper miRNAs were identified using heuristic cutoffs informed by simulation diagnostics.

Results

Classical AUC showed limited ability to detect improper profiles. DESeq2 remained robust for conventional differentially expressed features, but its recovery of improper profiles was lower under ranking-based screening. Across simulation settings, gAUC provided the most consistent recovery of improper signals, whereas LROC added useful complementary information under low-to-moderate overdispersion but degraded under extreme overdispersion. In the cervical cancer data, 416 miRNAs remained after preprocessing ROC-based screening flagged 56 candidate improper miRNAs, including 34 not prioritized by DESeq2. Several leading candidates were supported by published cervical cancer evidence.

Discussion

From a clinical biostatistics perspective, gAUC and LROC offer a pragmatic extension to standard RNA-Seq workflows by highlighting heterogeneous, non-monotonic expression patterns that may otherwise be overlooked. These indices should be interpreted as exploratory prioritization tools rather than confirmatory tests, because they do not provide formal error-rate control. Independent validation is required before clinical translation.

References

[1] doi: 10.1177/0962280214541095

[2] doi: 10.1007/s10182-020-00371-8

[3] doi: 10.1186/s13059-014-0550-8

[29-P7-03] OmicsBlendR: An Interactive R Shiny Application for Robust Biomarker Selection via Ensemble Machine Learning

Presenter: Hamdi Furkan Kepenek

Authors: Hamdi Furkan Kepenek [1]; Hatice Yağmur Zengin [1]

Affiliations: [1] Department of Biostatistics, Hacettepe University, Ankara

Introduction

Integrating high-dimensional genomic data with clinical phenotypes is essential for translational bioinformatics but is often hampered by single-algorithm biases and overfitting. We developed OmicsBlendR, an interactive R Shiny application designed to provide a robust, multi-algorithm machine learning environment for clinico-genomic biomarker discovery. We utilized an empirical Alzheimer's disease (AD) RNA-seq dataset to validate the framework's capacity to iteratively select and validate optimal predictive features.

Methods

OmicsBlendR automatically merges clinical datasets with transcriptomic matrices, offering bulk categorical relabeling and Log2(x+1) normalization. The core architecture utilizes an ensemble frequency-ranking pipeline to replace traditional single-model selection. Random Forest (Mean Decrease Gini), Support Vector Machines (Linear kernel weights), and XGBoost (Gradient boosting importance) are independently executed iteratively with random internal resampling. Features are ranked based on their consensus selection frequency. Selected biomarker panels are validated via Leave-One-Out Cross-Validation (LOOCV) or 5-Fold CV, reporting metrics including AUC (via pROC). Downstream modules map selected genes to clinical traits through Empirical Bayes differential expression analyses utilizing the limma framework (generating adjustable Volcano plots), Principal Component Analysis (PCA), and Spearman correlation networks via igraph and ggraph.

Results

Deployment on the AD dataset demonstrated that isolated algorithms yielded high selection variance. OmicsBlendR's iterative consensus approach successfully filtered transcriptomic noise, isolating a highly stable panel of candidate genes. Cross-validation confirmed robust predictive accuracy and high AUC scores. Furthermore, the limma-driven differential expression and interactive topological networks effectively visualized significant correlations between the optimally selected genomic signatures and underlying AD clinical parameters.

Discussion

OmicsBlendR provides a highly scalable, mathematically rigorous analytical framework for predictive bioinformatics. By integrating an iterative, frequency-based ensemble machine learning pipeline with advanced limma profiling, it significantly enhances the reproducibility of biomarker selection. Future developments aim to integrate automated Gene Ontology (GO) networks to further elucidate the biological significance of selected features.

[29-P7-04] A two-stage PD-first framework for response-modified transcriptomic inference in small, multi-arm longitudinal clinical datasets

Presenter: Kevin Vervier

Authors: Kevin Vervier [1]; Gunther Jansen [1]; Justin McMullen [1]; Nadia Solovieff [2]

Affiliations: [1] Novartis Pharma AG, Basel; [2] Novartis Pharma AG, Cambridge, MA

Introduction

High dimensional longitudinal analyses embedded in clinical trials are often considered underpowered, particularly when the question concerns whether pharmacodynamic (PD) molecular changes differ by clinical response. This reflects inherent sample-size constraints and analytical challenges related to multiplicity and interpretation. In multi-arm settings, unsupervised gene-wise testing across thousands of features and contrasts can rapidly exhaust statistical power - reducing the ability to detect true biological signals - and yield unstable significance patterns, increasing the risk of spurious mechanistic narratives driven by noise rather than biology. Practical analytical strategies are therefore needed to explicitly manage observer bias, multiplicity, and interpretation while remaining implementable using standard statistical tools familiar to applied clinical data science teams.

Methods

We analyzed whole-transcriptome data from 400 participants with paired baseline and end-of-treatment measurements in a randomized, placebo-controlled, three-arm study. Clinical response was defined as improvement in a disease activity score at end of treatment. The primary estimand was whether PD transcriptomic change differs by clinical response after accounting for overall PD effects.

Key analytical choices were aligned prior to execution using a lightweight pre-specification mechanism, including target estimands, primary contrasts, feature-selection strategy, and minimum effect-size thresholds.

A two-stage PD-first framework was applied. In Stage 1 (PD discovery), longitudinal PD effects were estimated using limma/voom, a linear-model framework that models the mean-variance relationship. PD feature selection relied on treat(), which tests contrasts against a pre-specified minimum log-fold-change rather than zero, prioritizing biologically meaningful PD effects. In Stage 2 (response-modified PD), hypothesis testing was restricted to PD-selected features. Because PD feature selection in Stage 1 is independent of clinical response, this substantially reduces multiple testing and

preserves power to detect true response-modified effects. Responder×arm×visit interaction models estimated response-associated deviation beyond PD, with interpretation guided by cross-arm effect-size.

Results

Restricting response-modified analyses to PD-selected features substantially reduced the multiple-testing burden and clarified interpretation when statistically significant findings appeared in one active regimen but not the other, despite highly concordant effect sizes.

Discussion

A PD-first, two-stage framework combined with lightweight pre-specification provides a practical approach for controlling power spending - thereby reducing false negatives and preserving sensitivity to true PD and response-modified effects - while improving interpretability in small longitudinal clinical studies.

[29-P7-05] Comparative Evaluation of Microbiome Simulation Methods Using Real Data

Presenter: Johan Verbeeck

Authors: maryam sh [1]; Olivier Thas [1]; Connie Musisi [1]; Johan Verbeeck [2]

Affiliations: [1] Hasselt University, Diepenbeek; [2] Hasselt University, Hasselt

Introduction

Recent advances in 16S rRNA sequencing have improved our understanding of the human microbiome and its links to disease, but analyzing microbiome data remains difficult due to sparsity, overdispersion, compositionality, and strong taxa correlations. Simulation frameworks are therefore essential, as they generate synthetic datasets with known properties, enabling controlled evaluation of analytical methods where real data lack clear benchmarks. Several approaches exist, including parametric, semiparametric, and nonparametric methods, each varying in how well they capture real data structures and ensure stability. In this study, we compare three widely used methods, MIDASim, SparseDOSSA2, and SPsimSeq, using a common template dataset to assess how well they reproduce key microbiome characteristics and to highlight their strengths and limitations for methodological benchmarking.

Methods

A template dataset from the Dietswap study was used, consisting of Day-2 gut microbiome samples from 37 individuals (130 taxa). Synthetic datasets were generated using four simulation methods, namely MIDASim (parametric and nonparametric), SparseDOSSA2, and SPsimSeq, each with differing assumptions about abundance distributions, sparsity, and correlation structures. Method performance was evaluated by comparing simulated and real data using taxon-level metrics, including mean abundance, variance, coefficient of variation (CV), and fraction of zeros, and by examining relationships between these metrics through scatter plots. To assess consistency, 100 independent datasets were generated (focusing on MIDASim). Taxa were grouped into abundance-based

quantile bins, and bin-level averages of variance, CV, and fraction of zeros were compared between simulated and real data. Differences were summarized using boxplots, where good performance is indicated by values close to zero with low variability.

Results

Comparison of the mean abundance versus variance plots across the four simulation methods (Figure 1) and the relationship between mean abundance and coefficient of variation (CV) (Figure 2a-d) showed that all methods reproduced the characteristics of the real data quite well. In contrast, the relationship between mean abundance and the fraction of zeros (Figure 2e-h) was better captured by nonparametric and semiparametric approaches. Figure 3 shows the differences between simulated and real bin-level means of log-variance, coefficient of variation (CV), and fraction of zeros across 100 parametric MIDASim simulations. The log-variance differences exhibit relatively small variability, indicating that parametric MIDASim framework consistently reproduced the variance patterns of the real data (Figure 3a). In contrast, the fraction of zeros showed greater variability across bins, suggesting that zero-inflation was not consistently captured (Figure 3b). Although the fraction of zeros was reasonably reproduced at the extremes of the abundance range, larger deviations were observed at intermediate abundances. The CV also showed deviations but was reproduced more consistently than the fraction of zeros (Figure 3c).

Discussion

Overall, the evaluated simulation methods reproduced key abundance-variance patterns of the template microbiome data, with nonparametric approaches better capturing sparsity patterns, while parametric MIDASim simulations showed strong consistency but were less accurate in modeling zero inflation.

[29-P7-06] Network Inference via Copula Models for Zero-Inflated Spatial Omics Data

Presenter: Giulia Capitoli

Authors: Giulia Capitoli [1]; Gaia Gubelli [1]; Vanna Denti [2]

Affiliations: [1] Bicocca Bioinformatics Biostatistics and Bioimaging Centre - B4, Department of Medicine and Surgery, University of Milano-Bicocca, Monza; [2] Proteomics and Metabolomics Unit, Department of Medicine and Surgery, University of Milano-Bicocca, Monza

Introduction

Biopsy-derived tissue samples analyzed through spatial omics technologies provide high-dimensional molecular measurements while preserving tissue spatial organization. These data exhibit strong spatial dependence and non-Gaussian distributions, including skewness and zero inflation. However, most classical statistical approaches neglect spatial structure and rely on restrictive assumptions, potentially biasing analyses.

This study addresses these limitations by developing and comparing methodologies that explicitly account for spatial effects and complex marginal distributions, with particular emphasis on copula-based modeling strategies.

Methods

Biopsy samples are represented as spatial grids, where each location is associated with multiple molecular signals. For each sample, molecular dependencies are encoded into networks, with nodes representing molecules and edges corresponding to pairwise associations.

Three approaches are considered to incorporate spatial information. A first strategy uses B-spline regression to remove large-scale spatial trends before estimating dependencies. A second approach adopts Spatial Autoregressive (SAR) models to account for local spatial autocorrelation. A third strategy relies on copula models to characterize joint dependencies independently of marginal distributions. Within this framework, a novel model combining copula functions with Zero-Adjusted Gamma (ZAGA) marginals is introduced to handle zero-inflated and skewed intensity distributions typical of spatial omics data.

Resulting networks are compared using graph-based distances, including spectral and edge-based measures.

Results

Simulation analyses indicate that accounting for spatial structure and non-Gaussian distributions substantially affects inferred dependency patterns, particularly in the presence of strong spatial autocorrelation or zero-inflated signals. Spatial preprocessing (B-splines and SAR) and copula-based approaches capture complementary aspects of the data, leading to networks with distinct topological properties.

When applied to biopsy samples of different thyroid lesion types, the framework reveals substantial inter-sample heterogeneity. A partial separation between broad pathological groups is observed, although no clear clustering emerges across all classes.

Discussion

From a statistical perspective, ignoring spatial dependence and marginal distributional features may lead to incomplete representations of molecular interactions. The proposed framework addresses these issues by combining spatial correction with flexible dependency models.

In particular, copula models with ZAGA marginals provide a principled approach to handle zero-inflated, non-Gaussian data within a network-based setting, to properly exploit the structure of spatial omics data.

[29-P7-07] A Data-Driven Geometric Metric for Long-Range Interactions in 3D Chromatin Structure

Presenter: Yunchae Jung

Authors: Yunchae Jung [1]; Insu Jang [2, 3]; Minsu Park [3]

Affiliations: [1] Chungnam National University, Dajeon; [2] Korea Research Institute of Bioscience and Biotechnology, Daejeon; [3] Chungnam National University, Daejeon

Introduction

Three-dimensional (3D) genome reconstruction from Hi-C data provides a geometric view of chromatin structure that complements traditional contact intensity analysis. In typical contact matrices, interactions between genomically adjacent loci dominate the signal, making it difficult to detect long-range interactions between bins that are distant along the genome but spatially close in reconstructed space. Identifying such distal pairs is essential for understanding large-scale chromatin organization, particularly in cancer genomics studies.

Methods

We introduce a data-driven geometric measure, formulated as a distance ratio metric, that summarizes the relationship between spatial distance in reconstructed three-dimensional space and linear genomic distance along the polymer. The proposed metric is scale-free and remains stable across multiple binning resolutions without relying on parametric decay functions. To evaluate its performance, we conducted simulation studies under diverse structural configurations and varying noise levels to assess its ability to identify non-trivial spatial proximities.

Results

Simulation results show that the distance ratio metric consistently highlights long-range interactions with unusually small spatial separation that are often masked by strong local contact patterns. When applied to cancer-associated Hi-C datasets, the metric reveals distinct patterns of large-scale chromatin structure that are not detectable through contact intensities alone. These findings indicate that the proposed approach captures complementary structural information beyond conventional contact-based analysis.

Discussion

The proposed distance ratio metric provides a statistically interpretable framework for studying chromatin structure in three-dimensional genome reconstruction. By integrating geometric reconstruction with genomic distance information, the method offers a robust tool for comparative analysis in cancer genomics and other contexts where long-range interactions play a critical role.

[29-P7-08] Adipose tissue k-means clusters show distinct patterns of targeted serum metabolites: the population-based KORA-MRI study

Presenter: Juliane Maushagen

Authors: Juliane Maushagen [1]; Elena Grune [2]; Christopher Schlett [1]; Lena S. Kiefer [3]; Karsten Suhre [4]; Jerzy Adamski [5]; Rui Wang-Sattler [6]; Annette Peters [7]; Fabian Bamberg [1]; Susanne Rospleszcz [2]

Affiliations: [1] Universitätsklinikum Freiburg, Klinik für Diagnostische und Interventionelle Radiologie, Freiburg; [2] Universitätsklinikum Freiburg, Klinik für Diagnostische und Interventionelle Radiologie, Freiburg im Breisgau; [3] Department of Diagnostic and Interventional Radiology, Eberhard Karls University of Tuebingen, Tuebingen; [4] Cornell University, Weill Cornell Medicine-Qatar, Doha; [5] Helmholtz Munich, Institut für Experimentelle Genetik, Neuherberg; [6] Helmholtz Munich, Institut of Translational Genomics, Neuherberg; [7] Helmholtz Munich, Institut of Epidemiology, Neuherberg

Introduction

Location and distribution of adipose tissue (AT) has more impact on cardiometabolic risk than overall body size. We have previously established five subphenotypes of body composition using k-means clustering of AT. Serum metabolite patterns of AT distribution could characterize potential pathophysiological pathways. We aim to identify metabolite signatures of single AT depots and body composition subphenotypes.

Methods

Targeted metabolites (Biocrates p180 kit) were measured in fasted serum for N=390 individuals from the population-based KORA-FF4 cohort (42% women, average age 56y). AT was measured by magnetic resonance imaging. Association of n=29 AT depots (visceral (VAT), subcutaneous (SAT), liver, pancreas, bone marrow, skeletal muscle, heart, kidney) and five subphenotypes with 146 metabolites and 40 derived indicators were investigated by linear regressions with confounder adjustment and multiple testing correction.

Results

Subphenoytpes were associated with 59, and single ATs with 275 metabolites/indicators, with VAT and SAT showing most associations. Compared to subphenotype I (low overall ATs), subphenotype II (average ATs) showed positive associations with sphingomyelin C20.2, hydroxysphingomyelin C22.1 and diacylPC with differently saturated C32 fatty acid. Subphenotype III (high muscle and bone marrow fat) was negatively associated with total lysophosphatidylcholines (lysoPCs) and total monounsaturated lysoPCs, while showing a positive association with total longchain acylcarnitines (C14-C18). Subphenotype IV (high SAT/VAT and liver fat) exhibited positive associations with short-chain acylcarnitines, alanine and aromatic amino acids. Subphenotype V (high pancreatic fat) was related to arginine and the ratio of ornithine and arginine as surrogate for ornithine synthesis. Three metabolites/indicators (lysoPC C 18.2, total polyunsaturated lysoPC, phospholipase A2 as ratio of lysoPC/diacyl and acylalkyl PC) were associated with all subphenotypes. These results were supported by the associations of individual ATs with metabolites.

Discussion

ATs and subphenotypes of body composition show distinct serum metabolite patterns, which can serve as a first step to characterize potential pathophysiological pathways.

[29-P7-09] LLM-supported translation of natural language statistical questions into model specifications and implementations in R and SAS for clinical research: Current status

Presenter: Jörg Sahlmann

Authors: Jörg Sahlmann [1]; Erika Graf [1]; Dominikus Stelzer [1]

Affiliations: [1] Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center - University of Freiburg, Freiburg im Breisgau, Freiburg im Breisgau

Introduction

Clinical researchers typically formulate analytical questions in natural language, while their implementation in R, SAS or other programming languages requires formal statistical syntax. The required translation is non-trivial and introduces a risk of errors.

Methods

This study investigates how Large Language Models (LLMs) recognize and implement the concepts of dependent variables, endpoints, covariates, interactions, random and fixed effects, and analytical objectives in statistical models.

Examples of clinical trial protocols are used as input for LLMs. The LLM is instructed to transfer the analyses for specific study endpoints planned in the protocol into statistical models and to implement the corresponding programming code. Additionally, free formulated descriptions in natural language are used as input.

Results

The recognition and implementation of the aforementioned concepts by LLMs is reported and evaluated. Differences between various programming languages are highlighted.

A finding relevant beyond the present context of LLM output assessment is that statistical models implemented in R and SAS may yield discrepant results, where the sources of divergence may be non-trivial to disentangle (e.g., defaults, numerical tolerances, and implementation details) and the available information does not always allow a definitive adjudication of which implementation reflects the intended estimand.

Discussion

Recommendations are given for model-specific, structured input prompts that remain as close as possible to natural language, but are also more likely to generate formally and substantively correct model specifications.

Wed 30.09.2026

P1 - Digital Health and Data Systems

Poster area: P1

[30-P1-01] Automatic Extraction of Clinically Relevant Parameters for Tumor Board Decision Support

Presenter: Imge Yüzüncüoglu

Authors: Imge Yüzüncüoglu [1]; Yuxuan Chen [2]; A. Altar Lüser [2]; Nikitha Shruthi Ramasetti [3]; Robert Oehring [3]; Felix Krenzien [3]; Philippe Thomas [4]; Sebastian Möller [2]; Roland Roller [5]

Affiliations: [1] German Research Center for Artificial Intelligence, Berlin; [2] German Research Center for Artificial Intelligence (DFKI), Berlin, Germany, Berlin; [3] Charité - Universitätsmedizin Berlin, Berlin; [4] German Research Center for Artificial Intelligence (DFKI), Berlin; [5] Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Berlin

Introduction Clinical decision-making relies on information documented in unstructured reports, which are often non-standardized and difficult to process efficiently. This challenge is particularly critical in multidisciplinary tumor board meetings, where decisions must be made under time constraints and based on incomplete information. Automated information extraction (IE) as presented in ADBoard supports structuring relevant clinical data for decision-making. Methods We developed and evaluated an IE pipeline for liver tumor cases, focusing on hepatocellular carcinoma, cholangiocarcinoma, and colorectal liver metastases. A dataset of 743 clinical documents from 134 patients was manually annotated to create ground truth labels for clinically relevant parameters. The task was formulated as named entity recognition (NER). We compared rule-based methods (regular expressions), transformer-based models, and an exploratory large language model (LLM) under realistic clinical constraints. Results Transformer-based models achieved the strongest overall performance, particularly for frequent parameters. RegEx performed competitively for well-defined patterns and proved robust in low-resource settings. The LLM showed inconsistent performance and did not outperform the other methods. Across all approaches, performance was also affected by data imbalance and annotation variability. Discussion Our findings indicate that current LLMs may not yet be reliable for specialized clinical IE under realistic constraints. Instead, a hybrid approach combining transformer-based and rule-based methods might be most effective. Furthermore, evaluation methodology substantially influences performance interpretation, as strict span-based metrics can underestimate clinically relevant results. These results highlight that improving annotation quality and evaluation design is as important as advances in model architecture for clinical IE.

The full paper pertaining to this abstract will be published in the GMDS journal "Medizinische Informatik, Biometrie und Epidemiologie (MIBE)."

[30-P1-02] An ICHD-3 Knowledge Graph for Structured Classification of Primary Headache Disorders: Development and Design

Presenter: Dorian Zwanzig

Authors: Dorian Zwanzig [1]

Affiliations: [1] Eberswalde University for Sustainable Development, Eberswalde

Introduction

Accurate classification of headache disorders according to ICHD-3 criteria remains a persistent challenge in primary care, with up to 50-60% of patients receiving incorrect diagnoses despite established guidelines [1]. The ICHD-3 taxonomy comprises over 200 diagnoses with complex hierarchical, temporal, and exclusionary criteria. Own preliminary benchmarking of open-source LLMs on this task identified retrieval quality - not model reasoning - as the primary bottleneck, with the performance gap decomposable into retrieval quality, noise, and intra-category differentiation.

Methods

ICHD-3 diagnostic criteria for primary headache disorders were formalized into a typed Neo4j graph schema encoding diagnostic hierarchies, temporal constraints, and attack frequency criteria. Diagnostic rules are represented as computable logical conditions, enabling deterministic traversal. The ontological foundation extends the HEADxACHE knowledge base [2]. A hybrid pipeline combining graph-based retrieval, LLM-assisted clinical entity extraction, and rule-based validation follows neuro-symbolic design principles [3]. A formal evaluation against the HeadAI benchmark is currently underway, using an embedding-based RAG as the baseline.

Results

Prior (submitted) benchmarking established the performance ceiling and gap structure: embedding-based RAG achieved approximately 57% top-1 accuracy on synthetic vignettes, dropping to 35-40% on real clinical cases. Under oracle conditions - where correct diagnostic criteria were provided directly in the prompt - the same open-source LLMs reached 92-96% accuracy, a gap of 37 percentage points attributable entirely to knowledge delivery.

The ICHD-3 graph schema for primary headache disorders has been fully implemented, covering attack duration ranges, associated symptom constellations, and computable exclusion and quantification criteria. Benchmark evaluation of the KG pipeline will be presented.

Discussion

Structured graph traversal is designed to approximate oracle retrieval conditions systematically, through deterministic criteria matching, explicit hierarchy encoding, and verifiable decision paths. The system can be integrated into the hospital's local infrastructure and enables GDPR-compliant decision support without the need for external data transfer. As the diagnostic reasoning layer relies on explicitly modeled knowledge rather than learned parameters, it remains applicable even where labeled clinical cases are scarce.

[1] Thiele A et al. Brain Sci. 202111(7):839.

[2] De Andrade JBC et al. BMC Neurol. 202525(1):489.

[3] Prenosil GA et al. Commun Med. 20255(1):491.

[30-P1-03] Building an Executable Knowledge Graph for Clinical Decision Support in Cardiology - Technical Case Report

Presenter: Emre Calik

Authors: Emre Calik [1]; Philipp Wiesenbach [2]; Christoph Dieterich [2]

Affiliations: [1] Universität Heidelberg, Heidelberg; [2] Medizinische Fakultät Heidelberg / Universitätsklinikum Heidelberg, Heidelberg

Introduction

State of the art Current LLM-based and vector-RAG-based systems support question answering over medical text, but they do not explicitly execute diagnostic pathways, inclusion/exclusion criteria, or multi-step recommendation logic. Knowledge-graph-based approaches improve explainability, but often require structured input and limited natural-language interaction.

Methods

Concept We developed a neuro-symbolic proof of concept that transforms selected parts of the 2024 ESC Guidelines for Chronic Coronary Syndromes [6] into an executable knowledge graph for deterministic, explainable, and source-traceable recommendation generation. The framework separates natural language understanding from clinical reasoning by restricting the LLM to entity extraction and delegating recommendation generation to an explicitly modeled rule graph.

Results

Implementation Selected guideline tables were manually translated into rule subgraphs in Neo4j, resulting in 47 executable rules covering 60 evaluable input constellations. Clinical concepts were mapped to SNOMED CT where possible. Free-text patient-like queries were processed with Qwen14b for entity extraction, followed by graph-based retrieval and deterministic rule evaluation. The framework was evaluated on 60 patient-like test cases and compared with a Qwen30b baseline using manually retrieved guideline tables as prompt context.

Discussion

Lessons learned The proposed KG-based system achieved 85.0% correct recommendations (51/60), compared with 46.7% (28/60) for the baseline, and produced no false recommendations, whereas the baseline produced three. Remaining KG errors were attributable to retrieval mismatches rather than failures of symbolic rule execution. These findings suggest that combining LLM-based query interpretation with explicit graph-based rule execution is a promising approach for reliable clinical decision support. The main remaining challenge lies in semantic retrieval and terminology alignment rather than downstream reasoning. References [6] C. Vrints et al.,"2024 ESC Guidelines for

the management of chronic coronary syndromes: Developed by the task force for the management of chronic coronary syndromes of the European Society of Cardiology (ESC) Endorsed by the European Association for Cardio-Thoracic Surgery (EACTS)," Eur. Heart J., vol. 45, no. 36, pp. 3415-3537, Sep. 2024, doi: 10.1093/eurheartj/ehae177.

[30-P1-04] From Data to Decision: A Coordinated AI Agent Workflow for Real-Time Patient Deterioration Prediction and Escalation Management

Presenter: Aman Deep

Authors: Aman Deep [1, 2]; Gernot Pucher [1, 2]; Kevin Kopp [2, 1]; Till Rostalski [2]; Felix Nensa [2]; Anna Stark [1, 2]; Jens Kleesiek [2]; Christian Reinhardt [1]; Christopher Sauer [2, 1]

Affiliations: [1] Klinik für Hämatologie und Stammzelltransplantation, Universitätsmedizin Essen, Essen; [2] Institut für künstliche Intelligenz in der Medizin (IKIM), Universitätsklinikum Essen, Universität Duisburg-Essen, Essen

Introduction

Clinical urgency assessment in acute care settings remains challenging due to the complexity of integrating real-time vital signs, lab results, and patient history while balancing clinical guidelines, patient preferences, and the risk of alert fatigue. Traditional systems often rely on static thresholds, leading to delayed interventions or unnecessary escalations. We aimed to develop and evaluate an agentic AI system that autonomously assesses patient urgency in real time within hematology and oncology wards, integrating clinical data, evidence-based guidelines, and patient-centered values to support timely, accurate, and ethical decision-making.

Methods

We designed a multi-agent architecture using OpenAI's agents SDK framework with Qwen-80B LLM, orchestrated by a Triage Agent that coordinates specialized sub-agents: A Patient Data Agent that extracts vital signs, lab values and clinical notes from our FHIR-based clinical data platform (SHIP), a Forecaster Agent for abnormal data, a Summary Agent, and an Urgency Monitor Agent that synthesizes findings into structured assessments. Smart de-escalation logic compares current with prior assessments using clinically meaningful thresholds to prevent over-alerting. The system is deployed in a development environment at UK Essen "WARD" (not for clinical use) as a dockerized microservices architecture using FastAPI, Langfuse, and PostgreSQL, ensuring full auditability, observability and scalability.

Results

Preliminary assessments confirm usability, with robust urgency assessments validated through retrospective analysis of SHIP EHR and qualitative clinician feedback.

Discussion

By enabling real-time, context-aware, and ethically grounded decision support, this agentic workflow potentially enhances clinical efficiency, supports early intervention, and strengthens patient-centered care, demonstrating the transformative potential of AI agents in acute care environments.

[30-P1-05] Security by Design: Medical Zero-trust AI Agents

Presenter: Betina Schnepf

Authors: Betina Schnepf [1]

Affiliations: [1] private, Heidelberg

Introduction

Artificial intelligence will fundamentally change many areas of social life and medicine. Particularly in medicine, it is becoming clear that the technology not only offers opportunities but also presents challenges. Algorithms can take over tasks that were previously performed by medical personnel. With appropriate training, they can generate suggestions and provide support.

This work addresses the question of safety in medical hospital software by improving Follow-up software with AI Agent Patterns and integrating them in a new Zero-trust AI agent architecture. We suggest a software solution that facilitates more powerful health care management with AI Zero-trust agent workflows.

Methods

With clinical follow-up studies group of patients are assembled and assessed after hospitalization. Determined by a factor of interest the subjects are contacted by phone with compiled questionnaires and the answers are collected and evaluated. We design and create AI agent workflows to assist with the assessment process. The AI workflow are based on security controls and uses speech-to-text, content extractor, and content validator agents. The agent responses are data interoperable with HL7 FHIR standard, ICD, and SNOMED classification. We implement the agents with Python and Rust.

Results

The advance secure AI agent workflow requires clear separation between control and execution. We improve and implement an multi-agent Zero-trust architecture with clear separation and execution, end-to-end observability across agents, consistent security at every boundary, and auditable, explainable agent decisions. Security controls are mTLS between agents, token scoped access, request validation and audit logging. The parser verifies, extracts content, and validates. Structured results are returned with an audit trail. The LLM driven orchestration uses plans and sub-agents, but also pauses for human approval.

Discussion

Clear boundaries enable security and observability with Zero-trust communication, identity-scoped-access, and policy enforcement at every agent boundary. Distributed tracing, token metrics, and CPU utilization monitoring across the entire agent chain monitors execution.

The new AI Architecture allows for easier follow-up support and assistance, better app and data security and hardware support. The AI workflow can be used to trigger and improve follow-up processes and actions. The new integrated agents improve the software and is work under progress.

[30-P1-06] From Model Performance to Real-World Impact: Why AI in Health Care Needs Emulated Intervention Designs

Presenter: Daniel Neumann

Authors: Daniel Neumann [1]; Tatjana Beppler [2]; Sebastian Stäubert [1]; André Scherag [2]; Markus Löffler [1]

Affiliations: [1] Institute of Medical Informatics, Statistics and Epidemiology, Leipzig; [2] Institute of Medical Statistics, Computer and Data Sciences (IMSID), Jena University Hospital, Jena

Introduction

Artificial intelligence, large language models (LLMs), and data-driven clinical decision support systems are moving rapidly from technical development into routine care. Yet their evaluation remains methodologically unresolved. Conventional prospective study documentation is often not sufficient to capture the real effects of adaptive digital interventions, because relevant mechanisms only emerge through implementation, workflow interaction, user behavior, and changing documentation practices. Conversely, retrospective model evaluation alone does not provide an adequate basis for causal effect estimation in real care settings. This is particularly relevant for AI-enabled systems, where technical performance and clinical impact are not equivalent. As a result, a substantial evidence gap persists between model validation, clinical utility, and regulatory acceptability. To introduce an emulated intervention design framework for the robust evaluation of AI-enabled clinical IT interventions under routine care conditions, while maintaining compatibility with the core principles of the Medical Device Regulation (MDR).

Methods

Within the German Medical Informatics Initiative, and specifically in the INTERPOLAR project, we developed a study design concept that addresses this gap. The framework is based on emulated intervention logic and combines principles from target trial emulation, causal inference, implementation science, and real-world evidence generation. It is intended for complex digital interventions such as AI-supported decision tools, LLM-assisted information processing, and workflow-integrated clinical IT. Rather than relying exclusively on either prospective data collection or retrospective model validation, the design introduces an intermediate methodological layer that jointly evaluates intervention logic, timing, context, use patterns, and data provenance.

Results

The proposed framework enables a more realistic and analytically rigorous estimation of IT effects than technical validation alone. At the same time, it avoids the excessive rigidity of purely prospective designs for highly adaptive systems. This is especially important for AI applications, where measurable effects depend on sociotechnical embedding, human-AI interaction, and downstream clinical decisions. The concept therefore creates a bridge between innovation and evidence generation without undermining MDR principles.

Discussion

Emulated intervention designs may become a key methodological foundation for the next generation of digital health evaluation. They provide a practical and scientifically robust interface between AI, clinical care, data infrastructures, and biometry - and may be essential if we want to generate credible evidence for digital interventions whose true effects only become visible in routine practice.

[30-P1-07] RetiGrade: A Client-Side, AI-Assisted Annotation Platform for Retinal Fundus Images with FHIR Questionnaire Integration and Explainable Image Classification

Presenter: Kemal Yildirim, Tobias Brix

Authors: Kemal Yildirim [1]; Dominik Heider [1]; Jens Storp [2]; Leo-Lion Eisfeld [2]; Nicole Eter [2]; Michael Storck [1]; Tobias Brix [1]

Affiliations: [1] Institute of Medical Informatics, University of Münster, Münster; [2] Department of Ophthalmology, University of Muenster Medical Center, Münster

Introduction

Systematic annotation of retinal fundus images is essential for ophthalmological diagnostics, clinical studies and AI model training. Existing tools like LabelMe or the Visual Geometry Group Image Annotator (VIA) offer generic annotation but lack domain-specific integration of ophthalmological measurement methods. Commercial solutions require cloud connectivity and proprietary formats. No current tool combines standardized medical questionnaires, explainable AI-assisted analysis and quantitative measurements while preserving data locality.

Methods

RetiGrade is a browser-based single-page application operating entirely client-side all data remains in the browser's IndexedDB and LocalStorage. It integrates three core components: (1) A questionnaire engine supporting HL7 FHIR Questionnaire resources with custom extensions for color-coded, grouped answer categories. (2) AI classification via a ConvNeXt model executed through Open Neural Network Exchange (ONNX) Runtime Web, detecting 19 diagnostic classes including diabetic retinopathy and glaucoma. For explainability, class-specific Gradient-weighted Class Activation Maps are rendered as semi-transparent overlays with fundus-adaptive circular masking and an interactive opacity slider. (3) Cup-to-Disc Ratio (CDR) measurement using fine-tuned MobileSAM:

an encoder produces image embeddings, two specialized decoders segment optic disc and cup. Binary masks are converted into editable polygons via edge detection and radial sorting the CDR is computed area-based using the Shoelace formula as sqrt(cupArea/discArea). Polygon vertices are stored as normalized relative coordinates (0-1). Additionally, semantic search via all-MiniLM-L6-v2 (Transformers.js) enables natural-language image filtering through cosine similarity. Special filters (e.g. Red-Free, Ben Graham, CLAHE) are implemented as SVG filter chains.

Results

RetiGrade processes retinal images entirely in-browser without server communication. The AI module classifies 19 classes with per-class explainability heatmaps. CDR segmentation produces interactive, editable polygons. A statistics dashboard visualizes the annotation progress, per-image time metrics and CDR distributions across three clinical categories (normal 0,5). Structured data export is available as CSV and JSON.

Discussion

RetiGrade demonstrates that browser-based AI inference, FHIR-standardized data capture and quantitative ophthalmological measurements can be unified in a single privacy-preserving research tool. XAI promotes transparency through visual decision localization, enabling researchers to evaluate model behavior. The client-side architecture ensures that no patient data leaves the local device, eliminating infrastructure requirements and simplifying deployment in clinical research settings.

[30-P1-08] Robust Skeleton-Based Physical Distress Detection via Illumination-Aware Graph Convolutional Networks

Presenter: Jiwon Im

Authors: Jiwon Im [1]; Minsu Park [1]

Affiliations: [1] Chungnam National University, Daejeon

Introduction

Ensuring patient safety in hospitals and care facilities requires robust monitoring systems capable of identifying physical distress under diverse conditions. However, conventional frameworks often fail in low-light indoor settings, as diminished illumination leads to inaccurate pose estimation and unreliable event classification. To address these limitations, this study proposes a framework for real-time physical distress detection that remains robust under adverse illumination. The proposed system combines skeleton-based motion analysis with scene-level context to maintain high detection accuracy where standard models typically degrade.

Methods

The framework extracts joint coordinates from video frames to represent body movements as skeletal graphs. A graph convolutional network is employed to encode spatial relationships between joints and capture temporal motion patterns, enabling the classification of routine daily actions versus physical distress indicators. To improve situational understanding, scene-level context features

are integrated to complement the skeletal representation. To mitigate performance degradation in low-light settings, an illumination-aware training strategy utilizes both standard and synthetically darkened images. This approach enhances the stability of joint extraction and ensures reliable classification without the need for computationally expensive image enhancement steps.

Results

Experimental results demonstrate that the proposed framework consistently distinguishes between routine actions and physical distress indicators across diverse lighting conditions. Notably, the classification performance under low-light conditions remains comparable to results achieved under normal illumination, showing high stability against lighting variations. The framework also achieves high-frequency processing speeds, confirming the feasibility of the system for real-time deployment in healthcare monitoring and indoor security scenarios.

Discussion

The proposed framework addresses a critical gap in automated monitoring by ensuring reliability in poorly lit indoor environments without compromising real-time performance. By integrating skeletal motion analysis with scene-level context, the system offers a robust solution for diverse clinical and security settings. These findings suggest that the framework can significantly enhance emergency response efficiency and patient safety, providing a scalable and practical tool for autonomous healthcare monitoring systems.

[30-P1-09] RadiAnnot - a pre-annotation tool for German radiology reports

Presenter: Martin Boeker

Authors: Claudio Benzoni [1]; Martin Boeker [2]

Affiliations: [1] Technical University of Munich, TUM School of Medicine and Health, Institute for AI and Informatics in Medicine (AIIM), Chair of Medical Informatics, TUM University Hospital, Munich; [2] Technical University of Munich, TUM School of Medicine and Health, Institute for AI and Informatics in Medicine (AIIM), Chair of Medical Informatics, TUM University Hospital, München

Introduction

Structured extraction of clinical concepts from German radiology reports remains challenging due to the limited availability of deterministic, auditable tools. Existing approaches often rely on proprietary backends or lack explicit assertion labelling (present/absent/possible). We present RadiAnnot, an open-source deterministic pipeline that transforms the Radlex ontology [1] entities into keyword engines for concept extraction, with integrated assertion labelling and clinicians review framework.

Methods

The pipeline is composed of four phases. First, ontology preparation transforms the RadLex OWL into tabular format using owlready2, extracting multilingual labels. Second, deterministic lexical engines are built with FlashText, using two tiers: strict and expanded. Third, mention extraction applies the engines to German radiology text, performing token-wise lemmatisation using spaCy or Stanza with

explicit span projection from lemma space back to the original text. Fourth, assertion labelling uses a Cotik/pyneg-style engine [2] with German-specific triggers: pre-negation (PREN), post-negation (POST), pre-uncertainty (PREP), post-uncertainty (POSP), pseudo-negation (PSEU),

and conjunction termination (CONJ). Scope resolution applies direction windows (forward for PREN/PREP, backwards for POST/POSP) with comma-sensitive expansion. Pseudo-negation suppression prevents false triggers like "no change" from incorrectly negating findings. Output formats are compatible with LabelStudio and INCEpTION. The review framework can route mentions to a general practitioner for high-volume verification and a senior neuroradiologist for escalated cases.

Results

The system produces mention-level extractions with provenance tracking, span coordinates (raw and lemma), tri-state assertion labels (PRESENT/ABSENT/POSSIBLE), confidence proxies based on trigger-entity token distance, and structured evidence trails for audit. Each mention includes RadLex class identifiers, German/English labels, and scan window metadata. The governance framework generates stratified metrics including inter-annotator agreement estimates, escalation yield, and model-versus-senior concordance.

Discussion

RadiAnnot provides the first open deterministic pipeline for German radiology concept extraction with full provenance and governance. The two-clinician framework addresses limited specialist availability through staged review. This enables clinically defensible annotation workflows for quality assurance, cohort identification, and decision-support applications

1 Langlotz, Curtis P. "RadLex: a new method for indexing online educational materials. " Radiographics 26.6 (2006): 1595-1597.

2 Cotik, Viviana, et al. "Negation detection in clinical reports written in German." Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016). 2016.

[30-P1-10] Machine Learning-based Detection of Biomedical Units in German Clinical Text Material

Presenter: Merle Spiegel

Authors: Richard Zowalla [1]; Martin Wiesner [2]

Affiliations: [1] Hochschule Heilbronn, Deutschland; [2] Hochschule Heilbronn, Heilbronn

Introduction

Clinical texts contain a large amount of information that is relevant for further treatment, diagnosis and medical research. Biomedical units provide important information about measurements and dosages. However, since most of the data is unstructured, accessing it is time-consuming and prone to errors. Precise recognition of measurements is therefore important when text data is processed.

The aim of this work was to develop machine learning models for the automatic recognition and extraction of biomedical units in clinical text.

Methods

The clinical text material was first pre-processed and automatically annotated by a program. Based on this training data, models were trained using various machine learning methods in Apache OpenNLP and Spacy. In OpenNLP, Maximum Entropy (MaxEnt), MaxEnt with Quasi Newton Optimization (QNO), and Perceptron were used. In Spacy, the models were trained using a transition-based approach for named entity recognition, with one model optimized for accuracy and one for efficiency. The models were then tested using a manually annotated evaluation dataset.

Results

In MaxEnt with QNO, Perceptron, and the transition-based approach in Spacy performed comparably on the evaluation dataset, each achieving an F1-score of 0.90. These approaches therefore appear to be similarly well suited for the recognition of biomedical units in clinical texts. MaxEnt without QNO performed worse, achieving a maximum F1-score of 0.71. All models recognized units very differently depending on the unit category.

Discussion

MaxEnt with QNO, Perceptron and the transition-based approach are suitable for recognizing biomedical unitsin clinical texts. Difficulties arise primarily from (i) the large number of varying spellings of units, (ii) the occurrence of unit symbols in other contexts and (iii) the limited amount of available clinical text material.

[30-P1-11] Clinical code mapping with LLM tool use: A Pilot for automated data extraction of medication and diagnosis information from unstructured clinical notes.

Presenter: Til Karsten Spreuer

Authors: Til Karsten Spreuer [1, 2]; Andreas Günther [3, 2]; Raphael W. Majeed [4, 2]

Affiliations: [1] Department of Internal Medicine, Universities of Giessen and Marburg Lung Center (UGMLC), Member of the German Center for Lung Research (DZL), Gießen; [2] Cardio-Pulmonary Institute (CPI), Institute for Lung Health (ILH), Gießen; [3] Department of Internal Medicine, Universities of Giessen and Marburg Lung Center (UGMLC), Member of the German Center for Lung Research (DZL), Giessen (Germany), Gießen; [4] Department of Internal Medicine, Universities of Giessen and Marburg Lung Center (UGMLC), Member of the German Center for Lung Research (DZL), Giessen

Introduction

Accurate clinical coding is fundamental to large-scale epidemiological studies, hospital billing, and the development of robust clinical decision support systems. Conventional methods for structured data extraction often rely on manual curation, which is prohibitively labor-intensive. Goal of this

project is to determine whether current state-of-the-art open-weight LLM models are suitable for extraction of structured data from non-English (German) clinical notes.

Methods

We anonymized 35 German doctor's notes of five patients from our hospital and developed one pipeline to extract and map medications and two for diagnoses. The latter compares a RAG based approach with an agentic AI. We ran these using three open-weight LLMs on a local GPU-PC.

Results

The F1 scores for diagnoses do not exceed 0.12. If we instead consider mapping to the broad category, then the F1 score increases to 0.18. For medications the F1 score is as high as 0.78 and even 0.95 if we consider trivial name extraction only.

Discussion

For trivial name extraction of medications, every encountered mistake is explainable. Due to limitations in the nature of the task it is infeasible to expect a perfect score of 1 in any of the coding scenarios. Further problems in LLM output and parsing are addressed.

LLMs excel at extraction. They are both suitable for information extraction from clinical notes for use in research databases. However, for a clinical setting where the treatment of patients would be dependent on LLM performance, the current state-of-the art open weight models are not accurate enough.

The full paper of this abstract will be published in the GMDS series "Studies in Health Technology and Informatics" with doi: 10.3233/SHTI260999, URL:https://dx.doi.org/10.3233/SHTI260999

[30-P1-12] HISTOPATH - Evaluating Open-Weight LLMs for Structured Annotation of Placental Pathology Reports

Presenter: Nanae Sai

Authors: Nanae Sai [1]; Hanna Elena Gatzemeier [1]; Christina Lohr [2]; Silke Große [1]; Anne Rüllich [1]; Tanja Groten [3]; Janine Zöllkau [1]; Yvonne Lindemann [1]

Affiliations: [1] Universitätsklinikum Jena, Jena; [2] IMISE - Uni Leipzig, Leipzig; [3] Universitätsklinikum Köln, Köln

Introduction

Histopathological examination of the placenta plays a central role in the clinical and scientific assessment of pregnancy-associated pathologies, such as fetal growth restriction (FGR) due to placental insufficiency. However, the comparability of histopathological free-text reports remains limited due to a lack of standardization. The HISTOPATH project evaluates the feasibility of using open-weight large language models (LLMs) to generate structured annotations from histopathological placental reports. We assess whether generic models without domain-specific medical training can reliably

extract clinically relevant findings and analyze the impact of model size and prompting strategies under resource-constrained conditions.

Methods

A total of 500 placental pathology reports are manually annotated using the INCEpTION platform, with 10% randomly undergoing monitoring to establish a gold standard. The annotation scheme is based on prior work within the research group and refined with expert pathological input.

Multiple open-weight LLMs of varying sizes (Gemma 3, Llama 3.1, DeepSeek R1, Phi-4, Mistral Small 3.1/4, and Mixtral) are evaluated using zero-shot, few-shot (five examples), and few-shot prompting with self-consistency. Performance is assessed using accuracy, precision, recall, and F1-score.

Results

Preliminary results on the dataset indicate that smaller models generally perform worse overall. Few-shot prompting and self-consistency lead to modest improvements for several models, although the effect varies depending on the model. Across models, macro F1-scores range from approximately 0.16 to 0.27, while baseline methods achieve substantially lower scores (0.04-0.05), indicating a clear improvement when using LLM-based approaches.

Discussion

Initial findings suggest that zero-shot prompting alone is not sufficient for reliable information extraction from histopathological free-text reports. Larger models achieve comparatively strong performance even in the zero-shot setting, while few-shot prompting can lead to further improvements for several models. Ongoing analyses, including the evaluation of self-consistency and post-processing of model outputs, will be presented at the conference. Overall, the results highlight both the potential and the current limitations of LLM-based approaches for scalable annotation in clinical text mining.

[30-P1-13] On-Premise Detection of a Guideline-Driven Oral Anticoagulation Shift in German Doctor's Letters using Local Large Language Models_

Presenter: Phillip Richter-Pechanski

Authors: Phillip Richter-Pechanski [1]; Christoph Dieterich [1]

Affiliations: [1] Klaus Tschira Institute for Integrative Computational Cardiology, Heidelberg

Introduction

A vast amount of German clinical routine data are still stored in unstructured doctor's letters. To make these data available for clinical research, medical information must be extracted from these letters entirely inside the hospital infrastructure, under strict data-protection and transparency constraints. We present an early clinical application of local large language models (LLMs) for medication information extraction (MIE) to detect a guideline-driven shift in oral anticoagulation treatment from

vitamin K antagonists (VKAs) to direct oral anticoagulants (DOACs) in German cardiology doctor's letters.

Methods

We analyzed two corpora from different time periods: 500 letters from CARDIO: DE (2020/21) and 538 routine letters from an internal 2012 corpus without gold annotation. For 2012, medication information was extracted automatically with an on-premise fine-tuned Llama-3.1-70b MIE pipeline for CARDIO: DE, gold annotations were used. Medication related to anticoagulants were identified via extracted medication mentions and medication-reason relations, normalized into VKA, DOAC, and OTHER classes using guideline-based lexicons. Furthermore, we analyzed the active-ingredient composition within the DOACs and performed a manual review of automatically annotated 2012 letters. We investigated whether a MIE pipeline trained on CARDIO: DE annotations can generalize to unseen older clinical letters and recover the expected treatment shift without additional manual annotations.

Results

DOAC proportion increased from 16.9% in 2012 to 59.9% in 2020/21, while VKA proportion decreased from 37.7% to 9.9%. Within the DOAC class, the dominant active ingredient shifted from rivaroxaban in 2012 to apixaban in 2020/21. Manual review showed that remaining errors were mainly linked to generic medication mentions and missing medication-reason relations rather than incorrect medication information extraction.

Discussion

We show that under strict clinical constraints, local LLM-based MIE models trained on CARDIO: DE generalize to unseen older German doctor's letters and can support medication trend analyses and provide a robust on-premise method for downstream clinical text analysis.

The full paper of this abstract will be published in the GMDS series "Studies in Health Technology and Informatics" with doi: 10.3233/SHTI260996, URL:https://dx.doi.org/10.3233/SHTI260996

[30-P1-14] Large Language Models for Clinical Note Simplification: A Systematic Review and Experimental Evaluation of Medical Text Readability

Presenter: Mara Teichmann

Authors: Mara Teichmann [1]; Pelin Özkara Menekseoglu [2]; Julian Schwarz [2]; Ute Karbach [3]; Anne-Maria Purohit [4]

Affiliations: [1] Technische Hochschule Brandenburg, Brandenburg; [2] Department of Psychiatry and Psychotherapy, Immanuel Klinik Rüdersdorf, Brandenburg Medical School Theodor Fontane, Brandenburg; [3] Institute of Medical Sociology, Health Services Research, and Rehabilitation Science (IMVR), Köln; [4] Technische Hochschule Brandenburg, Fachbereich Informatik und Medien, Brandenburg an der Havel

Introduction

Since the introduction of the Patient Rights Act, patients in Germany have gained legal access to their medical records, including clinical notes. However, these documents are typically written for healthcare professionals and are often difficult for patients to understand due to specialized terminology, abbreviations, and complex sentence structures. Large Language Models (LLMs) offer new opportunities to automatically simplify such texts while preserving medically relevant information.

Methods

This study investigates the potential of LLMs to improve the readability of German clinical notes by combining a systematic literature review with an experimental evaluation. Ten freely available LLMs were assessed using five synthetic clinical notes, which were simplified through standardized prompts designed to ensure linguistic clarity while maintaining content fidelity. Readability was analyzed using established indices (Flesch Reading Ease, Wiener Sachtextformel, LIX, SMOG, and Coleman-Liau) and complemented by a novel analysis of medical terminology and abbreviation density as indicators of domain-specific complexity.

Results

The results show that all LLMs substantially increased overall text length while consistently reducing the density of technical terms and abbreviations. However, no model achieved consistent improvements across all readability indices, highlighting limitations of traditional metrics in the medical domain. Models such as Mistral, ChatGPT, and Copilot demonstrated the highest efficiency in balancing linguistic simplification and text length.

Discussion

Overall, LLMs show strong potential to enhance the accessibility of clinical documentation for patients. However, their effectiveness depends on model selection, prompt design, and evaluation methodology. The findings suggest that conventional readability metrics should be extended with domain-specific measures to more accurately assess comprehensibility in medical texts.

[30-P1-15] WhisKI: A Local Speech-to-Text Tool with Speaker Attribution and LLM-Based Post-processing for Clinical Documentation and Research Data Management

Presenter: Christian Jäger

Authors: Christian Jäger [1]; Christoph Demus [1]; Patrick Liegel [1]; Ahmad Wattar [1]; Jan Christoph [1]

Affiliations: [1] Research Group (Bio-)Medical Data Science, Medical Faculty, Martin Luther University Halle-Wittenberg (MLU), Halle (Saale)

Introduction

Accurate documentation of spoken clinical interactions is essential for patient care and research but remains resource-intensive and often incomplete. While modern speech and language technologies offer effective support, their use in medical settings is frequently restricted by data protection requirements that prohibit cloud-based processing.

To address this gap, we developed WhisKI (Whisper+KI), a fully local application for speech transcription, speaker attribution, and structured postprocessing. In addition to enabling practical use, the system provides a framework for investigating emerging approaches to clinical speech intelligence.

Methods

WhisKI implements a modular, locally deployable pipeline combining Whisper-based speech recognition, speaker attribution, and locally hosted large language models (LLMs) within an interactive full-stack architecture (FastAPI backend, React frontend). Users can record or upload audio data, configure models, edit transcripts, and generate structured outputs such as summaries or extracted information.

Within this application context, we explored two research directions: (i) transcript-only LLM-based speaker attribution in comparison to audio-based attribution/diarization, and (ii) reference-free evaluation of LLM-based summarization using semantic similarity measures, stability analysis, and human assessment.

Results

The system enables fully local, end-to-end processing of conversational data and supports a range of clinical and research use cases, including physician-patient interactions and qualitative data collection. The interactive workflow allows user control and iterative refinement of results.

Our findings suggest that audio-based diarization remains the most robust approach for multi-speaker scenarios, while LLM-based attribution shows potential in structured clinical dialogues. For summarization, reference-free evaluation provides consistent signals for comparing model performance and supports the development of reliable postprocessing workflows.

Discussion

WhisKI demonstrates that privacy-preserving, locally deployable speech intelligence applications can bridge the gap between regulatory constraints and practical AI use in medicine. By combining a user-facing system with an extensible research framework, it enables both immediate application and systematic exploration of advanced methods for clinical documentation and research data management.

[1] A. Radford, J.W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, Robust Speech Recognition via Large-Scale Weak Supervision, arXiv. (2022). doi:10.48550/arxiv.2212.04356.

[2] A. Plaquet, and H. Bredin, Powerset multi-class cross entropy loss for neural speaker diarization, INTERSPEECH 2023. (2023) 3222-3226. doi:10.21437/interspeech.2023-205.

[30-P1-16] Local Speaker Attribution in German Medical Conversations: Comparing LLM Pipelines with Audio-Based Diarization

Presenter: Patrick Liegel

Authors: Patrick Liegel [1]; Jan Christoph [2]; Christoph Demus [3]; Ahmad Wattar [4]; Christian Jäger [4]

Affiliations: [1] Research Group (Bio-) Medical Data Science, Faculty of Medicine, Martin Luther University Halle-Wittenberg, Halle (Saale); [2] Martin-Luther-University Halle-Wittenberg, Halle; [3] Martin-Luther-Universität Halle-Wittenberg, Halle (Saale); [4] Research Group (Bio-) Medical Data Science, Faculty of Medicine, Martin Luther UniversityHalle-Wittenberg, Halle (Saale), Germany, Halle (Saale)

Introduction

Clinically useful conversation transcripts require accurate speaker attribution, yet privacy-sensitive medical settings often favor locally deployable processing. While audio-based diarization systems such as Pyannote are well established, large language models (LLMs) have mainly been explored as post-processing components in audio-guided pipelines. However, it remains unclear whether locally hosted LLMs can perform speaker attribution from transcript text alone, without audio access.

Methods

We conducted an exploratory shared-transcript evaluation on 32 German conversational recordings (16 two-speaker, 9 three-speaker, 7 four-speaker 97 minutes total), in which a single faster-whisper run defined the canonical transcript for all methods, thereby isolating speaker attribution from ASR variability. Two local LLMs, MedGemma-27B-IT and Llama-3.1-8B-Instruct, were evaluated using raw and presegmented transcript input and compared with Pyannote-based segment- and word-level mappings. All experiments were conducted under both known and unknown speaker-count conditions. The primary outcome was Word Speaker Error Rate (WSER) secondary outcomes were Validity Rate (VR), Real-Time Factor (RTF), and Speaker Count Accuracy (SCA).

Results

Pyannote-based pipelines achieved the lowest overall WSER and were clearly superior for three- and four-speaker recordings. In the two-speaker setting, MedGemma with presegmented input approached segment-level Pyannote performance (median WSER 0.10 vs. 0.06), and in a physician-patient subgroup achieved a slightly lower median WSER than Pyannote. Llama was consistently weaker, with lower VR and more critical failures. Segment- and word-level Pyannote mapping performed nearly identically, and all pipelines ran faster than real time on a single GPU.

Discussion

Transcript-only local LLM attribution is not a general substitute for audio-guided diarization, but MedGemma indicates a potential niche in short, strongly role-structured two-speaker medical dialogues. For broader multi-speaker use, audio-based Pyannote remains the more reliable choice, while

existing local LLM infrastructure for clinical documentation could be extended to speaker attribution with little additional effort where privacy constraints favor text-only, locally deployable processing.

[30-P1-17] Introducing diegr: Dynamic and Interactive EEG Graphics using R

Presenter: Stanislav Katina

Authors: Stanislav Katina [1, 2]; Zdeňka Geršlová [1, 3]; Martin Lamoš [4]

Affiliations: [1] Department of Mathematics and Statistics, Faculty of Science, Masaryk University, Brno; [2] School of Mathematics and Statistics, University of Glasgow, Glasgow; [3] Department of Mathematics and Physics, Faculty of Military Technology, University of Defence, Brno; [4] Brain and Mind Research Program, Central European Institute of Technology, Brno

Introduction

High-density electroencephalography (HD-EEG) is a non-invasive technique for measuring and visualizing human brain activity. During EEG data analysis, a wide range of graphical methods across multiple software platforms is typically employed. However, existing tools often lack the flexibility, scalability, and interactivity required by modern research and clinical applications. This poster introduces diegr, an R package developed to provide a dynamic and interactive visualization framework for HD-EEG data.

Methods

The package enables users to work directly with database tables, in addition to standard formats such as data frames or tibbles. This approach improves memory efficiency and facilitates the processing of large-scale HD-EEG datasets. diegr provides a suite of interactive visualization tools that support multiple stages of the analysis pipeline, including time-series visualization and topographical mapping. By leveraging the ggplot2 and plotly ecosystems, the package allows extensive customization of graphical outputs while maintaining interactivity.

Results

The functionality of diegr is demonstrated using HD-EEG recordings from a visual-motor task involving Parkinson's disease patients and healthy controls. The examples illustrate how the package can be used to explore and compare brain activity patterns, particularly within motor-related cortical regions.

Discussion

diegr is designed to capture the complex spatio-temporal structure of the HD-EEG data in a flexible and user-extensible framework. Compared to existing R-based EEG tools - which are often limited to specific tasks, data formats, or electrode configurations, or are no longer actively maintained - diegr offers a more general and modern solution. In addition, manufacturer-provided software typically restricts methodological transparency and graphical customization.

In contrast, diegr supports analysis of any properly structured dataset and allows users to extend workflows by integrating custom functions with its visualization tools. We believe that diegr represents a valuable addition to the current ecosystem of HD-EEG analysis tools.

References: GERŠLOVÁ, Zdeňka. diegr: Dynamic and Interactive EEG Graphics. 2026.

R package version 0.2.0. doi: 10.32614/CRAN.package.diegr

(development version: https://github.com/gerslovaz/diegr)

[30-P1-18] A Design Pattern Approach to User-Centered Visualization Concept Creation for Single-Patient Data Summaries in Clinical Decision Support

Presenter: Maria Zerlik

Authors: Maria Zerlik [1]; Ian-C. Jung [1]; Katharina Schuler [1]; Martin Sedlmayr [1]; Brita Sedlmayr [1]

Affiliations: [1] TU Dresden, Dresden

Introduction

Single-patient data visualizations support physicians in clinical decision-making, but designing effective summaries remains challenging. The user-centered design process is resource-intensive, existing design recommendations are often too general, and designers frequently rely on precedents. To address this gap, we developed a design pattern instrument to aid researchers and developers in creating user-centered single-patient data summaries, facilitating communication and enabling diverse designs through pattern combinations.

Methods

A scoping review informed an initial visualization technique map, which was evaluated in a focus group with medical informatics researchers and developers (N=7). Participants also rated potential additional components. A questionnaire captured their experience and ratings of instrument characteristics were, informing the development of a user group profile.

Based on these findings, an interactive prototype toolbox was iteratively developed by usability experts and evaluated in a workshop with medical informatics researchers and developers (N=7). Participants completed exploration tasks, rated the instrument, and provided feedback on strengths, weaknesses, and improvements.

Results

Participants highlighted the need for extensive filtering options in the visualization technique map and identified interaction, data abstraction, and layout patterns, including their interconnections, as key extensions. Practical examples and characteristics supporting collaboration and communication between developers and users were rated as highly important. The user group profile ("visualization concept creators") describes background, tool usage, expectations, concerns, and challenges in user-centered visualization design.

The final instrument includes 26 visualization technique patterns, browsable by function family, input data type, and visual shape, each with design reflection points and pattern combination options. A visualization process guide provides a workflow overview, knowledge base, and 30 additional combination patterns.

Overall instrument satisfaction was high (median 4.5 on a 5-point Likert scale). Participants valued the combination options and reflection points, while suggesting a clearer overview of the instrument.

Discussion

The instrument may facilitate more efficient development of user-centered visualization concepts for clinical decision support. Pattern profiles and combination options enable flexible, literature-informed decisions, while the process guide supports structured workflows. Future work will expand the pattern collection and evaluate the instrument with participants actively working on their own visualization projects. Despite the small sample, the multi-method, iterative process with experts strengthens the findings.

[30-P1-19] Beyond Black Boxes and Blank Slates: A Comparative Analysis of Human-in-the-Loop AI in E-PICOS vs. Traditional Statistical Software for Medical Education

Presenter: EMİNE Kanik, Ayse Canan Yazici

Authors: EMİNE Kanik [1]; Ayse Canan Yazici [2]

Affiliations: [1] AB Health Tech MedicReS, Houston; [2] izmir Bakircay University Faculty of Medicine, izmir

Introduction

The integration of artificial intelligence into statistical software presents a fundamental dichotomy: traditional tools (R, SAS, SPSS, Stata) function as 'blank slates' requiring extensive programming knowledge and manual configuration, while newer AI-assisted tools often operate as opaque 'black boxes' that bypass user understanding and control. Both extremes fail medical students and residents who must simultaneously develop statistical reasoning and produce standardized, error-free manuscripts. We evaluated the E-PICOS Multi-Table module, which introduces a novel 'human-in-the-loop' AI architecture, comparing its pedagogical and practical utility against traditional statistical software in medical education.

Methods

We conducted a comparative analysis of E-PICOS against SPSS, R, and Stata, focusing on the workflow from raw data to publication-ready baseline characteristics tables (Table 1). The evaluation specifically examined how each platform handles variable classification, descriptive statistics generation, and table standardization. We assessed the presence and effectiveness of controlled AI assistance - where the system suggests methodologies but requires explicit human confirmation - in supporting both learning outcomes and error prevention.

Results

Traditional software platforms demonstrate fundamentally different approaches to statistical analysis. SPSS and Stata require manual variable definition and separate commands for different data types, offering no active error prevention mechanisms. R provides high flexibility and transparency but demands steep learning curves, with external AI assistants (such as ChatGPT integrations) often generating code without teaching underlying statistical rationale or ensuring appropriate methodology. In contrast, the E-PICOS Multi-Table module implements a distinctive human-in-the-loop workflow: it automatically recognizes variable types (continuous, categorical, ordinal) from uploaded data, but mandates user verification before proceeding. The system then recommends appropriate descriptive statistics based on confirmed data types, requiring researcher approval before generating standardized, CONSORT/STROBE-compliant tables. This controlled automation prevents common statistical errors while simultaneously teaching appropriate variable classification and methodological reasoning.

Discussion

The E-PICOS Multi-Table module represents a paradigm shift from traditional statistical software by implementing human-in-the-loop AI architecture. Unlike SPSS or R, which assume prior statistical expertise, or fully automated AI tools that bypass learning, E-PICOS functions as an interactive tutor. By requiring human confirmation at critical decision points - variable mapping, test selection, and output validation - it ensures researchers maintain control and develop statistical literacy while benefiting from AI-driven efficiency and standardization. This balanced approach optimally supports both biostatistics education and the production of high-quality medical manuscripts, demonstrating how responsible AI integration can enhance rather than replace human expertise in clinical research.

[30-P1-20] How AI limits human thinking and what we can do about it - a theoretical concept

Presenter: Hannah Köpper

Authors: Urs Alexander Fichtner [1]; Hannah Köpper [2]; Susanne Weber [3]; Erik Farin-Glattacker [1]

Affiliations: [1] Section of Health Care Research and Rehabilitation Research, Institute of Medical Biometry and Statistics, Medical Faculty and Medical Center - University of Freiburg, Freiburg im Breisgau; [2] Albert-Ludwigs-University Freiburg, Medical Faculty, Office of the Dean of Academic Affairs, Freiburg im Breisgau; [3] Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center - University of Freiburg, Germany, Freiburg im Breisgau

Introduction

Large language models (LLMs) are increasingly integrated into academic work and higher education, fundamentally changing how information is accessed, processed, and communicated. While generative artificial intelligence (AI) offers new opportunities for knowledge acquisition and teaching, it also challenges established concepts of classic media literacy and critical information evaluation. Media literacy describes the ability to critically analyse, reflect on, and ethically evaluate media use.

With the rise of digital platforms, these concepts have been expanded to include social media literacy, but also AI literacy. This contribution discusses the competencies required to interact critically with LLMs in scientific and educational contexts.

Methods

The contribution presents a conceptual and theoretical analysis integrating perspectives from media and communication research with psychological models of information processing. Concepts such as media literacy, social media literacy, the Elaboration Likelihood Model, the Heuristic-Systematic Model, and Cognitive Load Theory are synthesized to examine how individuals evaluate and process AI-generated information, particularly in the context of (medical) research and higher education. Furthermore, existing AI literacy frameworks, including the AI Literacy Framework and the Dynamic AI Literacy Model, are reviewed and will be compared with a competency framework developed for higher education in medicine.

Results

The theoretical analysis suggests that LLMs simultaneously reduce extraneous cognitive load and increase the volume and accessibility of information. While this facilitates efficient information processing, it may also promote heuristic rather than systematic evaluation of content. Because LLMs can reproduce or simulate traditional credibility cues - such as academic language, structured arguments, or references - users may rely more strongly on superficial indicators of credibility. We describe this development as a potential "cue shift", in which established credibility signals are replaced or imitated by AI-generated cues. To address this challenge, a competency framework based on learning objectives is proposed consisting of three core domains: ethical-legal competence, technical competence, and practical application competence.

Discussion

Developing AI literacy is essential for enabling students, researchers, and educators to critically evaluate AI-generated outputs and use LLMs responsibly in research and teaching. Integrating structured competency frameworks into higher education curricula may help maintain scientific standards and foster reflective engagement with generative AI technologies.

[30-P1-21] Facilitating accessibility in statistical consulting

Presenter: Marianne Huebner

Authors: Marianne Huebner [1]; Michael Wallace [2]

Affiliations: [1] Michigan State University, East Lansing; [2] University of Waterloo, Waterloo

Introduction

Academic statistical consulting centers support diverse researchers, yet clients and consultants may face unrecognized communication and cognitive accessibility barriers.

Methods

This work proposes an organizational framework - Staff training, Processes and practices, Infrastructure, and Evaluation (SPIE) - to embed accessibility principles into consulting environments. Grounded in Universal Design for Learning (UDL), the framework promotes multiple modes of engagement, representation, action, and expression to improve the experiences of clients and consultants.

Results

Implementing accessible, inclusive practices ensures high quality statistical support, empowers all participants, and contributes to successful research outcomes. We provide examples which highlight strategies for structured communication, flexible meeting modalities, clarity in deliverables, and digital accessibility.

Discussion

Providing a respectful environment where differences are normalized and welcomed creates a sense of belonging for everyone. Applying these principles in statistical consulting centers can improve experiences for clients and consultants alike.

Reference

Huebner M, Wallace MP, Stukes F, Manski S, Montez-Rath ME. SPIE: An organizational framework to facilitate accessibility in academic statistical consulting centers. Journal of Clinical and Translational Sciences. 202610(1):e45. doi:10.1017/cts.2026.10712

[30-P1-22] Bridging Acceptance and Privacy in Pro-Adaptive Cognitive Assistive Technology: A Conceptual Research Model

Presenter: Robin Grashof

Authors: Robin Grashof [1]; Sarah Büscher [2]; Sinan Yavuz [1]; Annika Walker [1]; Thomas Plieger [1]; Bernhard Breil [1]

Affiliations: [1] Hochschule Niederrhein University of Applied Sciences, Krefeld; [2] Rhine-Waal University of Applied Sciences, Kamp-Lintfort

Introduction

Given the prevalence of neurodegenerative and other chronic diseases that manifest with cognitive symptoms - such as Parkinson disease, dementia, and attention-deficit/hyperactivity disorder (ADHD) - the impact on patients and their caregivers can be substantial. Consequently, the absence of a cure for these conditions remains a significant challenge. In consideration of the capabilities of cognitive assistive technologies to enhance daily living, there is a concern that these technologies may not adapt their functionalities to the unique course of individual diseases.

The advent of pro-adaptive cognitive assistive technologies (pro-CAT) aims to address this issue by leveraging AI to personalize support by processing numerous data sources and predicting disease

progression. In order to guarantee user-centered design in the context of AI for assistance with health data of partially seriously ill patients, an empirical survey is to be conducted to investigate relevant aspects of technology acceptance and perceived privacy. The objective of this study is to develop an economic psychological research model and analysis plan.

Methods

In light of high explained variance attributed to the Unified Theory of Acceptance and Use of Technology 2 (UTAUT2), the present study has focused on the most salient factors in the context of eHealth. To address privacy aspects, we considered a psychological model of health data disclosure, including trust in AI.

Results

The research model under consideration delineates the intention to use a pro-CAT as a consequence of the following factors: 1) Perceived Usefulness and Perceived Effort, 2) Perceived Privacy Risk, Perceived Data Control and Privacy Concerns, and 3) Trust in Automation, while taking account the moderating effects of further literature-based findings. The model will be evaluated through a cross-sectional survey combined with confirmatory statistical analysis. The participants will be given an introduction to the concept of pro-CATs and its implications for data privacy. This introduction will be in the form of a text and storyboard. Participants will then complete standardized items assessing the model's constructs.

Discussion

A literature-based research plan was developed to investigate technology acceptance and privacy concerns in pro-CAT. Data collection and analysis is remaining.

[30-P1-23] EPAMed: A Web-Based Patient Decision Aid for Structured Preference Clarification in Shared Decision Making

Presenter: Jakub Fusiak

Authors: Jakub Fusiak [1]; Andreas Wolkenstein [2]; Verena S Hoffmann [1]

Affiliations: [1] Ludwig-Maximilians-Universität München Institute for Medical Information Processing, Biometry, and Epidemiology (IBE), München; [2] Institute for Ethics, History and Theory of Medicine, Ludwig-Maximilians-Universität München, München

Introduction

Shared decision making requires patients to weigh benefits, risks, burdens, and practical implications across multiple criteria. Although patient decision aids support this process, many still lack a preference elicitation method that is both rigorous and usable in routine care [1,2]. We present EPAMed, a web-based patient decision aid for structured preference clarification within a broader shared decision-making workflow.

Methods

EPAMed was developed using Design Science Research with iterative input from patients, clinicians, epidemiologists, and medical ethics researchers. Implemented in R Shiny, the system combines pairwise trade-off questions, revisable answer histories, preference-profile visualization, ranked option presentation, and consultation preparation. Risk information is presented using natural frequencies, icon arrays, glossary support, and certainty labels. All processing occurs locally within the user session. The current implementation uses a pairwise multicriteria preference elicitation logic derived from PAPRIKA [3].

Results

EPAMed supports reflective preference construction by allowing users to answer trade-off questions, revise prior responses, inspect the resulting preference profile, optionally adjust weights manually, and compare ranked options in a structured interface. The system also generates exportable summaries for use in clinical encounters. A multiple sclerosis treatment decision was implemented as an illustrative instantiation, showing how the tool can be adapted to a new context by replacing the decision model and evidence tables while retaining the same interaction logic and interface components.

Discussion

EPAMed demonstrates how structured multicriteria preference elicitation can be operationalized within a transparent, user-centered patient decision aid. Its main contribution lies in linking preference clarification, evidence presentation, and consultation preparation in a single digital workflow. Beyond the presented example of multiple sclerosis, the architecture is intended to support transfer to other preference-sensitive clinical decisions.

References:

1. Stacey D, Lewis KB, Smith M, et al. Decision aids for people facing health treatment or screening decisions. Cochrane Database Syst Rev. 20241:CD001431. doi:10.1002/14651858.CD001431.pub6.

2. Witteman HO, Ndjaboue R, Vaisson G, et al. Clarifying Values: an updated and expanded systematic review and meta-analysis. Med Decis Making. 202141(7):801-820. doi:10.1177/0272989X211037946.

3. Hansen P, Ombler F. A new method for scoring additive multi-attribute value models using pairwise rankings of alternatives. J Multi-Criteria Decis Anal. 200815(3-4):87-107. doi:10.1002/mcda.428.

[30-P1-24] Patients Voices in Somnolink: Key Findings from an Advisory Board Meeting

Presenter: Srushhti Trivedi

Authors: Srushhti Trivedi [1, 2]; Silke Schicktanz [2]; Philip Zaschke [1]; Dagmar Krefting [1, 3]; Christoph Schöbel [4]; Thomas Penzel [5]

Affiliations: [1] University Medical Center Göttingen, Department of Medical Informatics, Germany, Göttingen; [2] University Medical Center Göttingen, Department of Medical Ethics and History of Medicine, Göttingen; [3] Georg-August-University Göttingen, Campus Institute Data Science, Göttingen; [4] Universitätsklinikum Essen, Essen; [5] Charité - Universitätsmedizin Berlin, Berlin

Introduction

Patients' self-help groups (SHG) in Obstructive Sleep Apnea (OSA) emerged in 1990s that aims to educate patients, improve research, clinical outcomes and discuss treatments. They publish materials and organize events. In Somnolink[1], the patient advisory board (PAB) aims at identifying key issues for patients in using sleep trackers and provide data for research.

Methods

A semi-structured, 2-hour PAB meeting was conducted at the DSGM Congress on November 29th in Hannover. This meeting was facilitated and moderated by Somnolink's team members to foster an open dialogue among seven patient members stemming from different SHG across Germany.

Results

The PAB consisted of seven members (five males and two females), representing a range of patient groups from both national and local levels across six German states: Baden-Württemberg, Saxony, Lower Saxony, Hessen, Schleswig-Holstein, and Bavaria. Patient representatives identified significant challenges regarding data protection, lack of standardization, and communication gaps between physicians and patients. They also emphasized the potential of using data to derive insights from parameters such as: hereditary factors or comorbidities like obesity for disease prediction. They expressed the need to have the ability to access, view, and understand health data that can foster transparency and informed decision-making. While patients do appreciate the potential of telemedicine such as remote monitoring and reduced wait times, but, barriers exist, including challenges for older patients and difficulties in finding compatible devices online. Concerns regarding data accuracy and reliability for wearables (such as smartwatches) were also raised, for e.g. Apple Watch's sleep tracking feature, may inaccurately indicate 'slept well' despite a poor sleep. SHG have been recognized as valuable, but their effectiveness is limited by inconsistent information, lack of gender representation, and despite efforts to increase accessibility, younger patient remains insufficient, leading to decline of many local groups.

Discussion

PAB discussions revealed significant issues of data protection, accuracy, reliability, reimbursement, access to support for new and younger patients along with the issues regarding wearables to track sleep and information on these new emerging technologies.

[1] Krefting D, et al. Somnolink- a national initiative to improve data sharing for better healthcare and research in OSA. GMS Med Inform Biom Epidemiol. 2025

[30-P1-25] Design and Evaluation of a Culturally Adapted mHealth Prototype for T2DM Management with Integration Potential into a National Health Information System

Presenter: Agatha Sanchez Delvo

Authors: Agatha Sanchez Delvo [1]; Andreas Schrader [1]

Affiliations: [1] Universität zu Lübeck, Lübeck

Introduction

In current digital health environments, a gap remains between clinical data management and patient-centred care. Many countries have implemented electronic health record systems that support the collection and management of clinical data, while mobile health (mHealth) applications supporting patient self-management are increasingly being developed. However, these components are often developed separately. This creates challenges in the management of conditions such as Type 2 Diabetes Mellitus (T2DM), where effective control depends on continuous patient engagement, behavioural support, and education. This highlights the need for patient-centred digital solutions that complement existing national healthcare infrastructures, while also being adapted to specific contexts and cultural factors. Using Costa Rica as a case study context, where the national electronic health record system (EDUS) is widely implemented, this study presents the design and evaluation of a culturally adapted mHealth prototype with integration potential into the national system.

Methods

A culturally adapted mHealth application for T2DM was developed following a Human-Centred Design (HCD) approach (ISO 9241-210). Requirements were derived from literature, stakeholder interviews, and context, user, and task analyses conducted in the Costa Rican healthcare setting. Based on these findings, a modular high-fidelity prototype was designed, including functionalities for education, nutrition, physical activity and medication support, as well as a conceptual approach to integration with EDUS. The prototype was iteratively refined through formative evaluation with patients and healthcare professionals and subsequently assessed in a summative evaluation. Usability was measured using the System Usability Scale (SUS).

Results

The evaluation showed very high usability, with a mean SUS score of 94.64. Participants highlighted the potential integration with the existing national system and the culturally adapted education module as key strengths. Both patients and healthcare professionals reported positive perceptions regarding usability and relevance for supporting self-management.

Discussion

The findings suggest that mHealth solutions aligned with national systems and local contexts can improve acceptance and usability. They also highlight the importance of designing digital health components, such as electronic health records and mHealth applications, as complementary rather than separate systems.

[30-P1-26] Telemetry-Based Staging of Usage Adherence in the Fitness Coaching Feature of the tala-med Cardio App

Presenter: George Jogho

Authors: George Jogho [1]; Nicole Wimmesberger [1]; Iris Geler [1]; Iris Tinsel [1]; Kathrin Helm [2]; Sarah Gerbach [2]; Isabelle Hempler [2]; Christian Plotzky [2]; Andy Maun [2]; Manuela Glattacker [1]

Affiliations: [1] Section of Health Care Research and Rehabilitation Research, Institute of Medical Biometry and Statistics, Medical Faculty and Medical Center - University of Freiburg, Freiburg im Breisgau; [2] Institute of General Practice/Family Medicin, Medical Faculty and Medical Center - University of Freiburg, Freiburg im Breisgau

Introduction

Routine app data can describe the usage of digital physical activity coaching beyond simple frequency indicators. This study applied the ABC (Ascertaining Barriers for Compliance) taxonomy to routine app telemetry. Originally defined for medication adherence, this taxonomy framework distinguishes three stages of adherence: initiation, implementation, and persistence. It was used here to quantify staged usage adherence to the tala-med Cardio fitness coaching feature and to illustrate the additional insight provided by this stage-specific approach.

Methods

A single-arm, prospective observational study was conducted with employees of a university medical center in Germany using a study-specific German-language Android Package Kit (APK) under routine conditions for four weeks. Telemetry to operationalize ABC adherence stages was captured via the web-analytics platform Matomo as time-stamped coaching-relevant events predefined by the developers (e.g., activity-planning edits, viewing weekly feedback). Initiation was defined as first login within 7 days after the activation code-dispatch date. Participant-specific weeks were defined as consecutive 7-day intervals starting individually with the first coaching-relevant event. For implementation, each week was classified as recommended active usage (events on ≥3 days), minimal recommended active usage (1-2 days), or no active usage (0 days). Weekly points (2/1/0) yielded an Implementation Score (0-8). Persistence was defined as activity in ≥3 consecutive study weeks.

Results

Routine app data can describe usage of digital physical activity coaching beyond simple frequency indicators. This study adapted the ABC (Ascertaining Barriers for Compliance) taxonomy, a three-stage medication-adherence framework (initiation, implementation, persistence), to quantify staged usage adherence to the tala-med Cardio fitness coaching feature using app telemetry and to illustrate the additional insight provided by this stage-specific approach.

Discussion

Overall, rapid initiation among activators was prevalent, whereas implementation and persistence were limited, with most participant-weeks showing no active usage. This work demonstrates that telemetry-derived adherence metrics building on the ABC medication-adherence framework can inform real-world stage-specific evaluation of digital interventions, provided event definitions are transparent, and telemetry capture is complete.

[30-P1-27] 'When will it finally be my turn?" - Development of the ambient gaming station 'TriNA" for medical knowledge transfer on triage in the emergency room

Presenter: Julia Holzlehner

Authors: Julia Holzlehner [1]; Maeve Querfurt [1]; Bente M. Sulimma [1]; Christopher Gross [1]; Alex Storjohann [1]; Anne Jacobsen [1]; Jolina Malakowsky [1]; Elias Gallizzi [1]; Susan Brauer [2]; Tino Hentschel [3]; Andreas Schrader [2]

Affiliations: [1] Universität zu Lübeck, Lübeck; [2] Institut für Telematik, Universität zu Lübeck, Lübeck; [3] Institut für menschzentrierte und interaktive Systeme, Universität zu Lübeck, Lübeck

Introduction

German emergency rooms are facing high numbers of patients, leading to increased risk of burnout among medical staff [1]. A considerable number of patients lack medical knowledge for self-assessment and understanding of organizational aspects [2]. A tangible user interface-based gaming-station named TriNA, a German acronym for "Triage in the Emergency Room", was developed, aiming to promote decision-making and health-literacy concerning triage. The overreaching goal is teaching users the emergency room procedures engagingly, and help them understand the decision-making processes on site.

Methods

The hard- and software of TriNA was developed in an interdisciplinary student project. This included a comprehensive literature review as well as an analysis, design, and evaluation phase covering the entire structure and flow of the game. The evaluation compromised of usability experts playing TriNA and answering a survey.

Data was analyzed using descriptive statistics.

Results

TriNA is a three-dimensional simplified emergency room board game with embedded displays and pieces with near field communication tags representing patients. To successfully play TriNA, users adopt the perspective of medical staff to make triage decisions and answer questions on first aid topics on embedded tablets to treat patients. In the evaluation phase, five participants tested the gaming-station in a laboratory environment, none of whom was professionally experienced with

triage. 80% reported a rather high or very high learning effect, while 100% would recommend playing TriNA. Participants liked playful learning and game design but criticized the fluency and the high number of moves required.

Discussion

Limited resources led to a small sample size reducing the generalizability of the results. Technical issues from the prototype may have affected user experience and data quality. Despite these limitations, the findings remain satisfactory. In particular, the high level of satisfaction suggests that game-based learning approaches hold considerable potential for effectively conveying complex clinical processes and enhancing procedural understanding.

[1] Naidoo R, Schoemann R. Burnout in emergency department staff: The prevalence and barriers to intervention. South African Journal of Psychiatry. 202329:2095. doi: https://doi.org/10.4102/sajpsychiatry.v29i0.2095

[2] Janerka C, Leslie GD, Gill FJ. Patient experience of emergency department triage: An integrative review. International Emergency Nursing. 202474:101456. doi: https://doi.org/10.1016/j.ienj.2024.101456

[30-P1-28] Health Literacy Game "Health Reality Check" - Development of an ambient gaming station for identifying medical fake news

Presenter: Vivian Litzcke, Julia Holzlehner

Authors: Vivian Litzcke [1]; Helke Abert [1]; Julia Holzlehner [1]; Aleyna Möller [1]; Christopher Gross [1]; Alex Storjohann [1]; Elias Gallizzi [1]; Nina Kopp [1]; Zeynep Vural [1]; Susan Brauer [2]; Tino Hentschel [3]; Andreas Schrader [2]

Affiliations: [1] Universität zu Lübeck, Lübeck; [2] Institut für Telematik, Universität zu Lübeck, Lübeck; [3] Institut für menschzentrierte und interaktive Systeme, Universität zu Lübeck, Lübeck

Introduction

Health information is increasingly being shared on social media, often without scientific evidence [1]. This makes it difficult for people to determine credibility. The interactive game-based learning station Health Reality Check (HRC) is designed to help adolescents to critically evaluate health information and identify misinformation by simulating a social media feed.

Methods

The development included a comprehensive literature review as well as an analysis, design, prototype and evaluation covering the structure, flow, hard- and software of the game. In the evaluation, adults (n=13, M=23.7 years) tested the station in groups. Observational data were analyzed using Mayring's content analysis [3], while questionnaire responses underwent descriptive analysis.

Results

The evaluation revealed an overall positive assessment of the learning station, especially the supplementary information of content contextualization and clarity of the video content. Observations showed strong interaction among players. Nevertheless, the evaluation identified usability issues such as readability and limited gameplay time. Perceived improvements in the ability to critically evaluate health information on social media were ambivalent.

Discussion

The results demonstrate a high level of engagement of participants, with the multiplayer mode fostering interaction and discussion, indicating potential to improve health literacy. A prospective and retrospective evaluation appears appropriate for future studies.

Overall, technical and pedagogical adjustments are necessary to achieve lasting improvements on the critical evaluation of health information on social media.

[1] Borges do Nasciment IJ, Pizarro AB, Almeida JM., Azzopardi-Muscat N, Goncalves MA, Björklund M, Novillo-Ortiz D., Infodemics and health misinformation: a systematic reviews of reviews. Bulletin of the World Health Organization. 2022. 100(9):544-561. doi: https://doi.org/10.2471/BLT.21.287654.

[2] DIN EN ISO 9241-210. Ergonomics of human-system interaction - Part 210: Human-centered design for interactive systems. Standard. Berlin: Deutsches Institut für Normung 2020 Mar.

[3] Mayring, P., Fenzl, T. Qualitative Inhaltsanalyse. In N. Baur & J. Blasius (Hrsg.). Handbuch Methoden der empirischen Sozialforschung. # 2. Wiesbaden: Springer Fachmedien 2019. p. 633-648.

[30-P1-29] Person-Centred Medical Advisory Assessments for Statutory Approval in Out-of-Hospital Intensive Care in Children 0-17 years of age

Presenter: Raphael Gruber

Authors: Raphael Gruber [1]; Ulfert Grimm [1]; Andreas Mappes [1]; Moritz Rau [1]; Thomas Gaertner [1]; Patrick Schunda [1]

Affiliations: [1] Medizinischer Dienst Hessen, Oberursel

Introduction

Since October 2020, German social law has enabled critically ill patients - primarily those who are tracheostomised and/or invasively ventilated - to receive out-of-hospital intensive care (OHIC) as a statutory health insurance benefit, supporting participatory living in home environments. Benefit approval requires a special medical prescription and a qualified assessment of weaning from invasive ventilation and/or decannulation potential by a personal, on-site assessment by the Medical Advisory Service.

Objectives

This analysis examines whether statutory requirements for OHIC benefit approval are being met, with focus on the performance and documentation of mandated weaning and/or decannulation assessments in children.

Methods

A retrospective secondary data analysis of on-site assessments performed by the independent Medical Advisory Service of Hessen, Germany, was conducted on children 0 thru 17 y. (years) of age, enrolled between July 1st, 2025 and February 28, 2026, entitled to statutory health benefits, having been prescribed OHIC and scheduled for benefit-relevant approval assessment at the place of care. All assessments were performed by trained tandem teams consisting of a specialist physician and a critical care nurse.

Results

We report analyses on 185 cases (male: 93 female: 92 median age: 7 y.) fulfilling the above criteria. 35 % (64/185) of the patients were tracheostomised with potential for decanulation in 3 cases 42 % (77/185) received continuous or intermittent invasive or non-invasive ventilation with potential for weaning in 7 cases. From a socio-medical standpoint, approximately 6 % of cases did not fulfil statutory prerequisites for benefit approval. Stratified by gender, pre-school (0-6 y.) and school age (7 to 17 y. included), all calculated proportions were comparable. The structured assessment approach will enable more differentiated, individualised evaluations over time.

Discussion

The findings indicate potential for further optimisation of systematic weaning and decannulation assessments also in children with OHIC. Mandatory, person-centred on-site evaluations by the Medical Advisory Service enhance OHIC awareness, transparency, appropriateness, and quality assurance in a high-cost care sector. Standardised assessments support sustainable development of statutory OHIC structures, warranting further large-scale retrospective evaluation.

[30-P1-30] Evaluating Patient Satisfaction in a Medical Center for Adults with Disabilities

Presenter: Marie Gerwek

Authors: Marie Gerwek [1, 2]; Lisa Schäfer [3]; Ulrike Holdgrün [1, 2]; Johanna Martin [3]; Toralf Kirsten [1, 2]

Affiliations: [1] Institute for Medical Informatics, Statistics and Epidemiology, Leipzig University, Leipzig; [2] Dept. Medical Data Science, Medical Informatics Center, University of Leipzig Medical Center, Leipzig; [3] University of Leipzig Medical Center, Leipzig

Introduction

Patient satisfaction is a significant indicator of healthcare quality and an essential component of quality management. Nevertheless, evaluating patient satisfaction among individuals with cognitive or neurological impairments poses specific methodological challenges. At the Medical Center for

Adults with Disabilities (MZEB) at the University of Leipzig Medical Center, no standardized approach previously existed to systematically collect patient feedback in the outpatient setting.

Methods

Two questionnaire versions were developed to address heterogeneous cognitive and communicative abilities. The plain language version targets patients with mild impairments or caregivers and assesses accessibility, communication, waiting times, infrastructure, and satisfaction with clinical specialties. An evaluation questionnaire was additionally designed and used to assess comprehensibility, length, and item design. The simple language version, designed for patients with more severe impairments, consists of eight concise questions with pictorial, smiley-based touchscreen responses and covers the same topics as the longer version in a simplified form. This version was reviewed by an external organisation specialising in accessibility and simple language.

Results

Both questionnaire versions were iteratively refined based on findings from the evaluation questionnaire and the external review. The plain language version was extended to comprehensively cover all relevant clinical specialties, while the simple language version was further optimized regarding wording, visual support, and usability.

Both instruments were successfully implemented in REDCap and piloted between November 2023 and September 2025. A total of 95 responses were collected for the plain language version and 53 for the simple language version. The respective completion rates were 84% and 100%.

Discussion

The adaptation of survey design to the requirements of disabled individuals has been demonstrated to facilitate a more inclusive evaluation of patient satisfaction. However, the implementation of such a system requires appropriate technological infrastructure and active support by healthcare staff. The lower completion rate of the plain language version indicates remaining barriers, necessitating further pilot testing. A subsequent pilot is planned to specifically assess feasibility regarding comprehension of questions and response options, technical implementation, and questionnaire length. After optimization, both instruments will be made openly available for reuse by external partners.

P2 - Prediction and Personalized Medicine

Poster area: P2

[30-P2-01] Predicting Event-Driven Analysis Timing in Clinical Trials: New Approach and Comparative Study

Presenter: Luzia Hanßum

Authors: Luzia Hanßum [1]; Anne-Laure Boulesteix [2]; Jakob Winkler [1]; Tim Müller [1]; Hannes Buchner [1]

Affiliations: [1] Staburo, München; [2] LMU Munich, Munich

Introduction

For time-to-event endpoints, clinical trial analyses are typically event-driven, meaning they are performed after a predefined number of events has been observed, rather than at fixed calendar times. Reliable estimation of the time required to reach this target is therefore essential for trial monitoring, operational planning, and efficient resource allocation. Such predictions are, however, challenging due to limited interim information and uncertainty in enrollment dynamics, event time distributions, and censoring mechanisms.

Methods

This work introduces in detail and systemically assesses a novel method, called TTAPred ('time-to-analysis prediction'), originally proposed by Kappeler (2025). It is tailored to predicting event-driven analysis timing based on observed cumulative enrollment and event data from early trial periods. Presented is a systematic comparison of the proposed method with two established approaches implemented in the R packages gestate and eventPred. The proposed method combines parametric time-to-event modeling with direct optimization to align predicted cumulative event counts with those observed over time. In contrast to existing approaches, TTAPred is designed to operate on aggregated life-table data rather than subject-level information. The predictive performance of the three methods is evaluated on two simulation designs, with differing enrollment schemes, event time distributions, and drop-out assumptions.

Results

Simulation results indicate that predictive uncertainty is highest at early interim analyses and decreases as additional data becomes available. No single method consistently outperforms the others across all simulation settings. Overall, TTAPred demonstrates competitive predictive performance while exhibiting robustness to modeling choices and reduced data requirements. The established approaches gestate and eventPred also perform well overall, but show reduced accuracy in heterogeneous settings with mixed event time distributions and, for gestate, at later evaluation time points.

Discussion

In conclusion, the results highlight that method performance depends strongly on the data-generating mechanism and the amount of available information. TTAPred provides a flexible and easily applicable alternative for event time prediction, particularly in settings where subject-level data is unavailable, while existing methods remain valuable tools within their more comprehensive trial planning and monitoring frameworks that encompass additional functionalities beyond event time prediction.

References

Kappeler, Christian (2025). Forecast of Events for Study Analysis Planning. Abstract, DAGStat 2025, Berlin. URL: https://hi.converia.de/custom/media/dag_stat_25_hi/DAGStat2025_BoA.pdf.

[30-P2-02] Prognostic model for predicting overall survival in patients with head and neck squamous cell carcinoma

Presenter: Jozefien Buyze

Authors: Jozefien Buyze [1]

Affiliations: [1] Johnson & Johnson, Beerse

Introduction

Prognostic scores for overall survival (OS) can be used in clinical trial design and analysis as inclusion criteria, stratification factors or adjustment covariates, or in clinical practice for risk stratification, patient counseling and treatment decision support. Here we develop a prognostic score for OS in patients with recurrent/metastatic head and neck squamous cell carcinoma (R/M HNSCC) and assess its performance.

Methods

Publicly available randomized controlled trial data was used to create and assess a prognostic score for OS for patients with R/M HNSCC. The score was built using a LASSO Cox PH model where the initial set of variables was: age, sex, tumor site, ECOG PS, prior radiotherapy, prior surgery, TNM stage, BMI, albumin, ALP, hemoglobin, LDH, white blood cell count, and platelet count. The model was developed using studies NCT00460265 and NCT00401323 downloaded from Project Data Sphere. Study NCT00415194 and the Flatiron Health Advanced Metastatic Head and Neck Squamous Cell Carcinoma database (https://flatiron.com/database-characterization) were used to assess the performance of the prognostic score and estimate the potential variance reduction if the score was used as an adjustment covariate.

Results

The prognostic score includes six variables: higher BMI and hemoglobin are associated with a lower risk of death and higher ECOG PS, white blood cell count, platelet count, and prior radiotherapy are associated with a higher risk of death. The estimated concordance statistic was 0.66 (95% confidence interval: 0.63 - 0.69) in NCT00415194 and 0.64 (95% CI: 0.62 - 0.66) in Flatiron database. The estimated variance reduction from adjusting for the score is 13.0% (95% CI: 7.4%-19.3%) in NCT00415194 and 8.5% (95% CI: 6.0% - 11.0%) in Flatiron database.

Discussion

This work shows how public data can be leveraged to develop and evaluate a prognostic score for OS in patients with R/M HNSCC. Applying this score as an adjustment covariate is projected to increase statistical power by 2 to 4 percentage points. The score showed good discriminatory power and calibration and can be utilized both in clinical practice, as a predictor of survival, and in future trials as an inclusion criterion, stratification factor or adjustment covariate.

[30-P2-03] Choosing the right kidney: advancing transplant decisions with a novel clinical prediction model

Presenter: Simon Schwab

Authors: Simon Schwab [1, 2]; Daniel Sidler [3]; Fadi Haidar [4]; Christian Kuhn [5]; Stefan Schaub [6]; Linard Hoessly [7]; Michael Koller [7]; Katell Mellac [7]; Agim Thaqi [8]; Ueli Stürzinger [9]; Bruno Tischhauser [9]; Andreas Elmer [1]; Nicola Franscini [1]; Nathalie Krügel [1]; Thomas Müller [1]; Lisa Straumann [1]; Isabelle Binet [5]; Déla Golshayan [10]; Kerstin Hübel [11]; Thomas Fehr [12]; Franz Immer [1]

Affiliations: [1] Swisstransplant, Bern; [2] University of Zurich, Zurich; [3] Inselspital, Bern University Hospital, Bern; [4] University Hospital of Geneva, Geneva; [5] HOCH Cantonal Hospital St. Gallen, St. Gallen; [6] University Hospital Basel, Basel; [7] Swiss Transplant Cohort Study, Basel; [8] Federal Office of Public Health (FOPH), Bern; [9] Patient Advisory Board, Swiss Transplant Cohort Study, Basel; [10] Lausanne University Hospital, Lausanne; [11] University Hospital Zurich, Zurich; [12] Cantonal Hospital Graubünden, Chur

Introduction

Despite the excellent overall prognosis of kidney transplantation, graft loss remains a serious concern. Currently, Switzerland lacks a widely accepted tool for personalized risk estimation. We therefore developed and evaluated a clinical prediction model to assess the risk of future graft loss at the time of organ offer.

Methods

KIDMO (kidney prediction model) was a retrospective multicenter national cohort study using data from the Swiss Transplant Cohort Study (STCS), collected at all six Swiss kidney transplant centers. Recipients underwent deceased-donor kidney transplantation between May 2008 and December 2021. The outcome was time from transplantation to kidney graft loss, with death as a competing risk, using a Fine-Gray subdistribution hazards model. A study protocol was registered with predefined candidate predictors and a statistical analysis plan [1].

Results

A total of 2,373 individuals with end-stage kidney disease underwent deceased-donor kidney transplantation. The median age at transplantation was 57 years (IQR 46-65) 1,500 recipients (63.2%) were male, 873 (36.8%) were female. 236 recipients experienced graft loss, and 336 died with functioning grafts. The cumulative incidence of graft loss was 4.4% (95% CI, 3.6%-5.3%) at 2 years and 8.7% (95% CI, 7.4%-10.0%) at 5 years. On internal validation, the model achieved a C-statistic of 0.65 (95% CI 0.60-0.71) at 2 years and 0.63 (95% CI 0.60-0.68) at 5 years, with good overall calibration. Internal-external cross-validation showed a 5-year C-statistic ranging from 0.58 to 0.72 across transplant centers with substantial between-center heterogeneity across all performance measures. Decision curve analysis demonstrated clinical utility, showing a higher net benefit than the established acceptance practice across clinically relevant decision thresholds.

Discussion

KIDMO provides individual risk predictions for graft loss at 2 and 5 years following transplantation, calculated at the time of organ offer. By offering both relative and absolute risk estimates, along with demonstrated clinical utility and net benefit, KIDMO enables evidence-based shared decision-making. The tool is available at data.swisstransplant.org/kidmo/.

References

1. Schwab S, Sidler D, Haidar F, et al. Clinical prediction model for prognosis in kidney transplant recipients (KIDMO): study protocol. Diagn Progn Res. 20237(1):6. doi:10.1186/s41512-022-00139-5

[30-P2-04] Development of Prediction Models for the Severity and Expression of Later Metabolic Disorders in Women with Gestational Diabetes Based on Clinical Findings During Pregnancy

Presenter: Yvonne Lindemann

Authors: Yvonne Lindemann [1]; Julia Palm [1]; André Scherag [2]; Friederike Weschenfelder [1]; Tanja Groten [3]

Affiliations: [1] Universitätsklinikum Jena, Jena; [2] Institut für Medizinische Statistik, Informatik und Datenwissenschaften, Universitätsklinikum Jena, Jena, Jena; [3] Universitätsklinikum Köln, Köln

Introduction

Gestational diabetes mellitus (GDM) is a significant risk factor for the development of prediabetes and type 2 diabetes mellitus (T2DM) in the years following pregnancy. Early identification of high-risk individuals provides an opportunity for targeted prevention strategies. This study aimed to analyze clinical parameters during or immediately after a GDM pregnancy to predict the long-term risk of developing T2DM or prediabetes.

Methods

The analysis was based on a retrospective cohort of women diagnosed with GDM between 1990 and 1998 at Jena University Hospital, supplemented by follow-up data from the prospective MuKiLOG study (2020-2022). Only the first documented pregnancy per patient was included. Long-term glycemic status (T2DM, prediabetes, or no diabetes) was determined from medical records up to 2024. Modeling was performed using multivariate imputation by chained equations (MICE) and penalized logistic regression with elastic net regularization. Receiver operating characteristic (ROC) analyses were used to determine optimal thresholds. External validation for T2DM was conducted using an independent GDM cohort from 2011 to 2017 at the same center. Separate models were developed for prediabetes.

Results

For T2DM prediction, 50 cases with and 65 without T2DM were included. Key predictors were maternal weight, glycemic values at GDM diagnosis, and cardiovascular parameters. The best-performing model (λ-min) achieved an AUC of 0.78 (vs. 0.63 for λ-1se), with a high negative predictive value (0.9) and an accuracy of 0.7 in external validation.

For prediabetes, 16 positive and 15 negative cases were analyzed. Important predictors included umbilical artery pH, cardiovascular variables, insulin therapy, and previous pregnancy complications.

Discussion

Predictive modeling shows promise for identifying women at high and low risk of developing T2DM or prediabetes directly after a GDM pregnancy. Umbilical artery pH emerged as a notable predictor. The high negative predictive value may support targeted reassurance of low-risk patients and more efficient allocation of follow-up resources. Despite limitations such as incomplete postpartum data, model performance was clinically relevant and comparable to established benchmarks. Further studies should focus on external validation and closing data gaps to enable clinical implementation.

[30-P2-05] Beyond Rule-Based Risk Assessment: Data-Driven Prediction of Pregnancy and Birth Complications in the FeMaR Project

Presenter: Kerstin Rubarth

Authors: Kerstin Rubarth [1]; Jon Barrenetxea [2]; Elias Grünewald [3]; Barbara Tabernig [4]; Julia Sommer [4]; Stephanie Roll [5]; Felix Balzer [3]; Brigitte Strizek [6]; Thomas Reimhold [7]; Eva Jansen [8]; Florian Recker [9]; Silke Wegener [4]

Affiliations: [1] Charité - Universitätsmedizin Berlin, Institut für Biometrie und klinische Epidemiologie, Berlin; [2] Charité Universitätsmedizin Berlin, Institute of Medical Informatics, Berlin; [3] Charité - Universitätsmedizin Berlin, Berlin; [4] Charité - Universitätsmedizin Berlin, Institute of Medical Informatics, Berlin; [5] Institute for Biostatistics and Informatics in Medicine and Ageing Research, University Medical Centre Rostock, Rostock; [6] Department of Obstetrics and Prenatal Medicine, Center for Obstetrics and Gynecology, University Hospital Bonn, Bonn; [7] Charité - Universitätsmedizin Berlin, Institute of Social Medicine, Epidemiology and Health Economics, Berlin; [8] Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Institute of Medical Sociology and Rehabilitation Science, Charitéplatz 1, 10117, Berlin; [9] Department of Obstetrics, University of Cologne, Köln

Introduction

Pregnancy and birth complications remain a major challenge in obstetric care. In Germany, risk assessment is traditionally guided by the Mutterpass (maternity record), which is based on predefined risk factors and remains largely unchanged since 1986. This approach lacks individualization and may classify a large proportion of women as high-risk, limiting clinical usefulness.

The FeMaR (Fetomaternaler Risiko-Score) project aims to develop and validate multivariable prediction models for a broad range of maternal and neonatal complications to enable more precise and individualized risk stratification.

Methods

A large retrospective cohort of over 110,000 pregnancies derived from routinely collected clinical data at two tertiary care university hospitals was analyzed. Multiple clinically relevant outcomes

were considered, including hypertensive, metabolic, infectious, and severe maternal and neonatal complications.

Machine learning models based on gradient boosting were developed using a nested cross-validation framework to ensure robust performance estimation and to prevent overfitting. Separate models were trained using baseline variables available early in pregnancy and extended feature sets including longitudinal information.

Model performance was assessed using the area under the receiver operating characteristic curve (AUC), precision-recall AUC (PR-AUC), and log-loss. Calibration was evaluated on independent test folds using calibration-in-the-large and calibration slope.

For clinical interpretability, a three-level risk stratification system ("traffic light") was derived based on sensitivity- and specificity-driven thresholds.

Results

Discrimination ranged from AUC 0.62 to 0.81 across outcomes, with several models achieving good performance (AUC $\geq$0.70), particularly when incorporating longitudinal information. Models incorporating longitudinal information outperformed baseline-only models. Calibration was acceptable, although slight miscalibration was observed for rare outcomes.

The proposed traffic light system enabled clinically interpretable stratification into low-, intermediate-, and high-risk groups. Compared to traditional rule-based classification approaches, the models enable more differentiated risk allocation, reducing the proportion of individuals classified as high-risk while maintaining clinically relevant sensitivity.

Discussion

The results highlight the potential of data-driven approaches to complement and modernize existing risk assessment strategies in obstetrics. By moving beyond static systems, such as the Mutterpass, the FeMaR models enable individualized, dynamic risk prediction and more targeted care. Recruitment for prospective external validation in a real-world cohort has recently been initiated.

[30-P2-06] Beyond viral load: a prediction framework for vertical HIV transmission risk in low-resource settings

Presenter: Sandra Rjeschni

Authors: Sandra Rjeschni [1]; Kira Elsbernd [1]; Bindiya Meggi [2]; Anange F. Lwilla [3]; Ilesh V. Jani [2]; Issa Sabi [3]; Noemi Castelletti [1, 4, 5]; Arne Kroidl [1]

Affiliations: [1] Institute of Infectious Diseases and Tropical Medicine, LMU University Hospital, LMU Munich, Munich; [2] Instituto Nacional da Saúde, Marracuene; [3] National Institute for Medical Research, Mbeya Medical Research Center, Mbeya; [4] Fraunhofer Institute for Translational Medicine and Pharmacology ITMP, Immunology, Infection and Pandemic Research, Munich; [5] Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI), University Medical Center, Johannes Gutenberg University, Mainz

Introduction

Vertical HIV transmission (VHT) remains a major global challenge, with 120,000 children acquiring HIV in 2024. WHO's high-risk classification relies heavily on recent maternal viral load (VL), which is often unavailable in low-resource settings, underestimating transmission risk. We developed and validated a two-model prediction framework, one incorporating VL and one without, to ensure robust risk stratification regardless of VL availability.

Methods

Both models use weighted logistic regression (WLR) with elastic net regularization, with VHT cases upweighted during training. The framework was trained on the LIFE study (2019-2023 N=6580 137 transmissions), with performance evaluated on a held-out test set (30%), and validated in a temporally and operationally distinct cohort (LIFE2Scale 2024-2026 N=3834 41 transmissions). We simulated varying VL availability (0-100%) by proportionally blending predictions from both models and compared performance against WHO criteria. All modelling steps were repeated across 30 stratified seeds metrics represent seed-averaged means. Weighted random forests, decision trees and support vector machines served as comparators.

Results

WLR demonstrated higher sensitivity than WHO across all VL availability levels below 100%. In the LIFE test set, sensitivity at 0% VL was 0.74 vs. WHO at 0.14 with 100% VL, sensitivity was comparable (0.89 vs. 0.90). Specificity ranged from 0.61-0.79 across VL availability levels. In LIFE2Scale, sensitivity at 0% VL was 0.62 vs. WHO at 0.07 with 100% VL, results remained comparable (0.75 vs. 0.76). Specificity ranged from 0.70-0.88. (AUROC: LIFE 0.89/0.76, LIFE2Scale 0.86/0.70, 100%/0% VL). Precision was low (0.02-0.08) due to severe class imbalance since WLR outputs predicted probabilities, the decision threshold can be adjusted to clinical priorities. All comparators performed similarly or worse.

Discussion

By combining two complementary models, the framework maintains strong sensitivity regardless of whether maternal VL is used, offering flexible deployment across settings with varying VL availability. The sensitivity drop in LIFE2Scale reflects a fixed threshold that can be recalibrated based on health system priorities and budgeting, with the corresponding specificity gain supporting generalizability. This approach offers a robust addition to WHO's high-risk classification, addressing a critical gap in pediatric HIV prevention in low-resource settings.

[30-P2-07] Machine Learning for HIV Prediction in Complex Survey Data

Presenter: Farzana Osman

Authors: Farzana Osman [1, 2]; Mohanad Mohammed Adam Mohammed [2]; Temesgen Zewotir [2]; Nonhlanhla Yende-Zuma [3]

Affiliations: [1] Centre for the AIDs Programme of Research (CAPRISA), Durban; [2] Discipline of Statistics, School of Agriculture and Science, University of KwaZulu-Natal, Durban; [3] South African Medical Research Council, Durban

Introduction

Supervised machine learning (ML) offers powerful tools for risk prediction, yet most applications to survey-based data ignore complex survey design features, including sampling weights and clustering. Ignoring these features may produce biased estimates and misidentify key risk factors. This study evaluated whether incorporating complex survey design into ML framework affects HIV prediction performance, calibration, and variable importance.

Methods

We analysed pooled data from two waves of the HIV Incidence Provincial Surveillance System (HIPSS n=20 048), conducted in uMgungundlovu District, KwaZulu-Natal, South Africa. Twelve models were evaluated across three algorithms: logistic regression, random forest, and XGBoost. Two cross validation strategies (standard 10-fold and enumeration area-based group cross-validation) and two weighting approaches (unweighted and probability-weighted). Data were split 80/20 into training and test sets. Performance was measured using AUC, sensitivity, specificity, and F1 score. Calibration was assessed using mean absolute calibration error across predicted probability deciles. Variable importance was compared using normalized permutation importance scores.

Results

Weighted HIV prevalence was 35.8% (95% CI: [34.7 - 36.8%), higher among women (44.6% 95% CI: 43.3 - 45.9%) than men (26.3% 95% CI: 24.9 - 27.7%). XGBoost achieved the highest discrimination across all model configurations, with AUC values ranging from 0.835 to 0.837 compared to random forest (AUC: 0.828-0.832) and logistic regression (AUC: 0.830-0.832). Sampling weights and clustering-aware cross-validation had minimal impact on discrimination and calibration. Variable importance differed meaningfully across algorithms. Within logistic regression, weighting had minimal effect, with age groups 30-44 years and TB testing history remaining primary predictors. Within tree-based models, survey weighting elevated adolescents aged 15-19 years as the highest priority risk group. Tree-based algorithms additionally identified sexual behaviour variables, including absence of condom use at first sex and forced sexual debut, not captured by logistic regression.

Discussion

ML algorithms demonstrated good discriminatory performance and calibration for HIV prediction using survey data. Survey weighting meaningfully altered variable importance rankings redirecting attention from older adults towards adolescents aged 15 to 19 years and revealing additional sexual behaviour risk factors. Although ML techniques are applicable to complex survey data, survey weighting is essential for valid identification of population-level HIV risk factors to inform targeted prevention strategies.

[30-P2-08] Integrating Machine Learning and Logistic Regression for Stroke Risk Prediction in India's Ageing Population

Presenter: Anita Pal, Vineet Kumar Kamal

Authors: Anita Pal [1]; Vineet Kumar Kamal [2]

Affiliations: [1] INCLEN Trust International, New Delhi; [2] All India Institute of Medical Sciences (AIIMS), Kalyani

Introduction

India's rapidly ageing population, projected to reach 319 million aged 60+ by 2050, faces increased stroke risk due to comorbidities. Existing Western models may not suit India's diversity. This study aimed to develop and validate a stroke risk prediction model for adults aged ≥45 years and compare statistical and machine learning approaches.

Methods

Using LASI Wave 1 data (65,395 participants), stroke was identified through self-reports. Predictors included demographic, behavioural, clinical, and family history variables. Random Forest enabled feature selection, followed by logistic regression. Model performance was assessed via AUC and calibration, with leave-one-state-out cross-validation and external validation using Delhi data.

Results

Among the 65,395 participants, 1,195 (1.83%) reported having a stroke. Random Forest was used for feature selection to identify the most important predictors, which were subsequently incorporated into a logistic regression model. The final model showed that increasing age and higher body mass index (BMI) were significantly associated with higher stroke risk. Low physical activity, particularly among individuals reporting no activity, markedly increased the likelihood of stroke. Clinical conditions such as hypertension, diabetes, and heart disease emerged as strong predictors, while family history of stroke, hypertension, and diabetes further elevated risk. Females were associated with a comparatively lower risk than males.

The model demonstrated good discrimination, with an Area Under the Curve (AUC) of 0.79 (95% CI: 0.63-0.95) in the external validation dataset. Internal-external cross-validation across states showed consistent performance, supporting generalizability across diverse regions of India. Calibration analysis indicated moderate agreement between predicted and observed outcomes, with slight overfitting and some underestimation of risk. While machine learning models showed comparable performance, logistic regression was preferred for interpretability.

Discussion

This study demonstrates an effective, context-specific stroke prediction model for India's ageing population, identifying key modifiable risks such as physical inactivity, hypertension, and BMI. The model showed consistent performance across regions. Despite limitations like low prevalence and cross-sectional design, it provides a practical tool for early risk identification and prevention strategies.

[30-P2-09] Comparing Logistic Regression and Machine-Learning Models for Predicting Psychiatric Hospitalization and Developing a Web-Based Decision-Support Prototype

Presenter: Ebru Turgal

Authors: Ebru Turgal [1]; Hüseyin Şehit Burhan [2]; Furkan Bıyıklı [2]; Ali Uçar [2]; Pınar Özültürk [2]

Affiliations: [1] Ankara University Faculty of Medicine Department of Biostatistics, Ankara; [2] Department of Psychiatry, Basaksehir Cam and Sakura City Hospital, İstanbul

Introduction

Early prediction of psychiatric hospitalization can improve patient safety and support timely clinical intervention. This study aimed to (i) identify predictors associated with psychiatric hospitalization and (ii) compare the performance of traditional logistic regression with modern machine-learning models, while examining the impact of different train-test splits (70/30 vs 80/20) and decision-threshold optimization.

Methods

Psychiatric hospitalization (yes/no) was modeled using sociodemographic, clinical, and laboratory predictors. Variable selection was performed using LASSO logistic regression and selected features were entered into a multivariable logistic model. Additional supervised models (Random Forest and XGBoost) were trained using stratified splitting (70/30 and 80/20). Performance metrics included AUC, accuracy, sensitivity, and specificity. Optimal classification thresholds were derived using the Youden Index to balance sensitivity and specificity.

Results

The multivariable logistic regression model showed good discrimination (AUC = 0.79) and identified older age, previous suicide attempt, prior psychiatric diagnosis, and previous admission as independent risk factors, whereas female sex was protective.

Under the 70/30 split, AUCs ranged 0.68-0.72 across models. After threshold optimization, logistic regression achieved sensitivity 0.72 with balanced specificity (0.66).

Under the 80/20 split, logistic regression and LASSO improved further (AUCs 0.789-0.792). With optimized thresholds, LASSO achieved accuracy 0.74, sensitivity 0.85, and specificity 0.72, providing the most clinically balanced performance. Weighted XGBoost yielded higher sensitivity but substantially reduced specificity, limiting screening usefulness.

Discussion

Both traditional and machine-learning models demonstrated clinically meaningful predictive performance however, interpretability and stability favored logistic regression and penalized regression. The comparison of 70/30 versus 80/20 splits indicated modest yet consistent performance gains with the larger training set. Threshold tuning substantially changed clinical operating characteristics, underscoring the trade-off between missing high-risk patients and generating false alarms. Given the clinical priority to minimize missed hospitalizations, LASSO provided the most practical balance.

A prototype web-based decision-support tool was developed to support clinical translation, and future work will focus on external validation and implementation studies.

[30-P2-10] Predicting treatment challenges due to fentanyl use among people receiving medication-assisted therapy for opioid use disorder: A machine learning approach

Presenter: Myanca Rodrigues

Authors: Myanca Rodrigues [1]; Alannah McEvoy [2]; Tanvi Punjani [3]; Zainab Samaan [1]

Affiliations: [1] Queen's University, Kingston; [2] McMaster University, Hamilton; [3] The Research Institute of St. Joe's Hamilton, Hamilton

Introduction

The increasing prevalence of fentanyl in North America's illicit opioid supply has fundamentally altered the clinical course of opioid use disorder (OUD), with higher risks of overdose and treatment instability. Among people receiving medication-assisted therapy (MAT), fentanyl use may pose unique challenges, including continued opioid use and elevated overdose risk. However, tools to identify patients at highest risk in this evolving context remain limited. We aimed to develop machine learning (ML) models to predict key treatment challenges among people receiving MAT, with a focus on fentanyl use.

Methods

We conducted a population-based cohort study among people receiving MAT for OUD in Ontario between 2011 and 2021. Data from two prospective cohort studies of people with OUD across 54 Canadian Addiction Treatment Centres in Ontario were merged. Candidate predictors included demographic characteristics, clinical history, and indicators of fentanyl use. Machine learning algorithms, including random forest and gradient boosting (XGBoost), are being developed to predict two binary outcomes: (1) overdose and (2) continued opioid use during treatment. Model performance is being evaluated using measures of discrimination (area under the receiver operating characteristic curve [AUC]) and calibration, with internal validation.

Results

The training sample included 3,486 people receiving MAT, with a mean age of 40.0 years (standard deviation 11.0) and 1,916 males (55.0%). A total of 128 (3.7%) had documented fentanyl use, while 758 (21.7%) experienced an overdose and 2,000 (57.4%) demonstrated continued opioid use during follow-up. Emerging findings suggest that fentanyl use, in combination with younger age and male sex, is associated with increased risk of overdose, whereas fentanyl use alongside a higher burden of comorbid psychiatric illness is more strongly predictive of continued opioid use. Preliminary model performance demonstrates moderate discrimination (AUC ~0.65-0.70) with good calibration across risk strata, with ongoing refinement and validation underway.

Discussion

Fentanyl use appears to result in differential risk profiles for treatment outcomes among people receiving MAT. These findings highlight the importance of incorporating fentanyl use into predictive models of MAT outcomes. ML approaches can capture complex interactions between fentanyl use and patient characteristics, enabling more precise identification of people at risk of specific treatment challenges.

[30-P2-11] Development and Validation of a Complex Chronic Conditions Algorithm Based on Claim Data in Pediatric Patients with Disabilities

Presenter: Taemi Youk

Introduction

Comorbidity indices developed for adults, such as the Charlson Comorbidity Index (CCI), do not adequately capture the spectrum of chronic conditions in pediatric populations. Children and adolescents, particularly those with disabilities, often present with heterogeneous and complex health needs that differ substantially from adult populations. The Feinstein-based Complex Chronic Conditions (CCC) framework has been proposed as a pediatric-specific alternative however, its applicability to children with disabilities remains insufficiently explored, particularly when applied to administrative claims data in different healthcare systems. This study aimed to implement a claims-based CCC algorithm using Korean National Health Insurance data and to evaluate its predictive performance for mortality compared with the CCI in pediatric patients with disabilities.

Methods

We conducted a nationwide retrospective cohort study using Korean National Health Insurance claims data. A total of 157,071 pediatric patients (<19 years) with registered disabilities between 2006 and 2022 were included. The CCC algorithm was adapted to Korean claims data using diagnostic and procedure codes. Mortality risk was estimated using Cox proportional hazards models. Model discrimination was assessed using the C-statistic, with 95% confidence intervals obtained via bootstrapping. Comparative analyses with the CCI were performed, along with subgroup analyses by age, sex, and disability type.

Results

The CCC algorithm identified a broader range of clinically relevant chronic conditions than the CCI. Patients classified under CCC showed higher mortality risk. The CCC-based model demonstrated strong discrimination for mortality, with a C-statistic of 0.833 (95% CI, 0.826-0.840), outperforming the CCI-based model, which had a C-statistic of 0.714 (95% CI, 0.703-0.722). These findings were consistent across subgroup analyses.

Discussion

We implemented and validated a claims-based CCC algorithm for pediatric patients with disabilities in Korea. Compared with the adult-oriented CCI, the CCC demonstrated superior predictive performance and better captured clinically relevant conditions in this population. While CCC was originally developed for general pediatric populations, our findings suggest its potential utility in children with disabilities. This approach provides a useful framework for risk stratification and health services research using administrative data. Further studies are warranted to assess broader applicability across healthcare settings.

[30-P2-12] Enhancing Mortality Risk Prediction: The Role of Wearable-derived Physical Activity Metrics

Presenter: Yuhe Wang

Authors: Yuhe Wang [1]; Cameron Razieh [2]; Alex V. Rowlands [3]; Tom Yates [3]

Affiliations: [1] University of Leicester, Leicester; [2] Leicester Real World Evidence Unit, Diabetes Research Centre, Leicester General Hospital, University of Leicester, Leicester, UK, Leicester; [3] Diabetes Research Centre, University of Leicester, Leicester, UK, Leicester

Introduction

Traditional risk prediction models predominantly rely on demographic, clinical and some behavioural factors, e.g., smoking and body mass index (BMI), but fail to capture the dynamic nature of lifestyle behaviours. There is growing recognition of the importance of physical function and fitness metrics in predicting health outcomes. For example, self-reported walking pace is strongly associated with mortality and may serve as an alternative to clinical markers for risk prediction. However, traditional models struggle to incorporate continuous physical activity data, which wearable devices can now provide in real time. This study aims to evaluate whether adding to or replacing traditional biomarkers with wearable derived metrics can improve the prediction of mortality risk.

Methods

Using UK Biobank data, participants were stratified by health status. The base model included seven traditional factors: age, sex, smoking, BMI, systolic blood pressure (SBP), total cholesterol-to-HDL ratio (CHR), and deprivation. 13 accelerometer metrics, including daily activity volume (AvACC24h), intensity gradient (IG), and peak 10-minute activity (M10CONT), and 1 combined indictor of physical activity volume and intensity (AvACC24h+IG) were evaluated. Analysis followed three stages: (1) adding factors, (2) substituting SBP or CHR with accelerometer metrics, and (3) replacing both SBP and CHR. Model performance was assessed using Harrell's C-index and Net Reclassification Index(NRI).

Results

The largest improvements were seen in individuals with health conditions, where M10CONT and the combined indicator (AvACC24h + IG) consistently outperformed the traditional clinical model. Replacing SBP and CHR with the combined indicator improved the C-index by 0.026 (95%CI: 0.014, 0.031) and NRI by 13.4% (7.6% to 19.3%). M10CONT, as the strongest individual predictor, improved the C-index by 0.024 (0.016, 0.033) and NRI by 16.4% (8.7%, 24.0%). Both M10CONT and the combined indicator surpassed the previous strongest self-reported indicator, walking pace.

Discussion

These findings show that non-invasive, real-time movement metrics from wearable devices can match or even exceed the predictive power of traditional clinical measurements and self-reported indicators like walking pace. Our findings also suggest that considering accelerometer data reduces misclassification of risk by capturing a comprehensive picture of body function and health.

[30-P2-13] An Ensemble Bayesian Model Averaging Framework to Combine Deep Learning Models for 30-Day ICU Readmission Prediction

Presenter: Emanuele Koumantakis

Authors: Emanuele Koumantakis [1, 2]; Konstantina Remoundou [3]; Ioanna Roussaki [3]; Alessia Visconti [1]; Paola Berchialla [1]

Affiliations: [1] Department of Clinical and Biological Sciences, University of Turin, Turin; [2] CPO Piemonte, Città della Salute e della Scienza di Torino University Hospital, Turin; [3] Institute of Communication and Computer Systems, National Technical University of Athens, Athens

Introduction

Intensive Care Unit (ICU) readmissions increase morbidity, mortality, and healthcare costs. Existing deep learning (DL) models for predicting 30-day ICU readmission show limited clinical utility due to poor calibration and overconfident predictions that can mislead clinicians. To address these limitations, we developed an Ensemble Bayesian Model Averaging (EBMA) -based framework to integrate predictions from multiple DL, machine learning (ML), and traditional scoring models. While EBMA rewards discriminative power, it also penalizes models that provide overconfident yet inaccurate predictions, resulting in more reliable and clinically meaningful risk estimates.

Methods

We analyzed first ICU admissions for adult patients (≥18 years) from the MIMIC-III database (n=32,784 stays). We used as input for our EBMA-based framework predictions from 17 DL and 6 ML models as well as the traditional SWIFT and APACHE-II scores. We randomly split the data into training (80%), calibration (13%), and testing (7%) set. To assess calibration and discrimination of the obtained ensemble, we used Brier Score and AUROC, respectively. We compared these results against those of the individual models.

Results

Our EBMA-based framework achieved a Brier Score of 0.051 and an AUROC of 0.728, outperforming all individual models in calibration while matching or exceeding their discrimination performance. These results remained robust after excluding lower-weighted models and, notably, even after removing the highest-weighted and best-calibrated model.

Compared with individual models, similarly well-calibrated models (i.e., with similar Brier Scores) achieved AUROC values of at most 0.683, indicating inferior discrimination. In contrast, the most discriminative model (AUROC = 0.716) exhibited poor calibration (Brier Score 0.148). While our EBMA-based framework obtained only a marginal improvement over this model (+1.7% AUROC, likely due to limited model diversity), by substantially improving calibration it yielded more reliable and clinically valuable predictions.

Discussion

Well-calibrated predictions are more valuable for clinical decision-making than high discrimination alone. Our EBMA-based framework reliably integrates multiple models, delivering superior calibration while maintaining a discrimination comparable to the best-performing individual model. Our framework shifts away from the "winner-takes-all" paradigm of standalone models toward a more robust, ensemble-based strategy. Future work should focus on external validation and on boosting model diversity to improve discriminative performance further.

[30-P2-14] Comparing Semi-Automated Surveillance Algorithms for Surgical Site Infections

Presenter: Leander Perz

Authors: Leander Perz [1, 2]; Robin Ristl [1]; Sneschana Neschkova [2]; Elisabeth Presterl [2]

Affiliations: [1] Medical University of Vienna, Center for Medical Data Science, Institute of Medical Statistics, Vienna; [2] Medical University of Vienna, Department of Hospital Epidemiology and Infection Control, Vienna

Introduction

Currently surveillance of surgical site infections (SSIs) requires manual review of patient data, which makes continuous monitoring feasible only for selected surgeries. Semi-automated surveillance aims to preselect patients for review who seem likely to have had an infection, extending monitoring capacity. This study compares five binary prediction algorithms ranging from classical approaches to modern transformer models. We aim to establish a model to identify SSI after coronary artery bypass graft (CABG) surgery and to investigate whether advanced models provide actual benefits.

Methods

We evaluated logistic regression, CART, random forest, XGBoost, and the transformer-based model TabPFN on data from 5545 CABG surgeries performed at the University Hospital Vienna, where SSIs were identified through manual chart review. Logistic regression required multiple imputation

while other models handled missing data natively. To control for overfitting, we used 10-fold cross-validation and calculated out-of-fold predictions. As a performance metric, we computed cross-validated AUC with the R package cvAUC.

The design of semi-automated surveillance requires very low false negative rates (FNR) since patients not preselected will not be manually reviewed. To better understand the effectiveness of the models, we compared the true negative rates (TNR) at an FNR of at most 5%. To compute this, we used threshold averaging to create a single ROC curve per algorithm.

Results

CART had the lowest AUC of 0.65 (95% CI: 0.59-0.71) while logistic regression and XGBoost performed similarly, both achieving 0.73 (0.69-0.77). The highest AUC was achieved by TabPFN, 0.77 (0.73-0.81), followed closely by random forest, 0.76 (0.72-0.80). The results were similar when looking at the TNR at the specified FNR. The TNR was highest for TabPFN at 0.372, followed by random forest (0.337) and logistic regression (0.336).

Discussion

TabPFN provided the best overall discrimination, but simpler models offer greater interpretability. In particular, CART produces a decision tree which can be qualitatively evaluated and avoids the implementation of opaque models. Models such as logistic regression provide a middle ground with solid performance at a low FNR.

The next steps are developing validation strategies to enable real-world implementation and extending the covariate space to match performance requirements for clinical use.

[30-P2-15] EHR-Based Identification of Polyneuropathy and Myopathy Patients

Presenter: Elisabeth Félicité Nyoungui

Authors: Elisabeth Félicité Nyoungui [1]; Shamim Ahmed [2]; Nicolai Truong [2]; Hanna Wedemeyer [3]; Verena Schuster [4]; Rudolf Mayer [4]; Dagmar Krefting [1]; Jana Zschüntzsch [3]; Richard Röttger [5]

Affiliations: [1] University Medical Center Göttingen, Department of Medical Informatics, Göttingen; [2] University of Southern Denmark, Department of Mathematics and Computer Science, Odense; [3] University Medical Center Göttingen, Department of Neurology, Göttingen; [4] SBA Research gGmbH, Vienna; [5] University of Southern Denmark, Department of Mathematics and Computer Science, Odense M

Introduction

Polyneuropathies (PN) are disorders that damage the peripheral nerves responsible for movement and sensation, leading to muscle weakness, numbness, tingling, and limb pain. Myopathies (MY) are primary muscle diseases that also cause weakness and muscle atrophy. Because both conditions show overlapping clinical signs, relying on history and physical examination alone rarely yields a definitive diagnosis. Applying machine learning to electronic health record data (EHR) offers new

opportunities to more efficiently distinguish between PN and MY. The aim of the study was to develop and evaluate machine learning models using EHR data to reliably differentiate polyneuropathy from myopathy, paving the way for improved diagnostic support tools.

Methods

The analytic dataset comprised 2,181 adult patients (PN = 1,853 MY = 328) and 3,701 features, which were extracted from the Data Integration Center (DIC) of the University Medical Center Göttingen. This data routinely harmonizes EHRs in compliance with the German Medical Informatics Initiative. The features were selected based on recommendations from physicians, the literature, and statistical analysis. We implemented Logistic Regression, Random Forest, and XGBoost models, their hyperparameters optimized with Grid Search, and addressed class imbalance using SMOTE.

Results

Random Forest and XGBoost models achieved robust discrimination (F1 Macro scores of 0.82-0.84 and AUC-ROC scores of 0.92-0.93) when trained on demographic data, feature-engineered variables, laboratory test results, and ICD-10 codes. This confirms that even relatively low-detailed electronic health records, when carefully maintained, contain sufficient information to support early differentiation between PN and MY in routine clinical practice.

Discussion

Guided feature selection ensures plausibility by integrating neuromuscular expert knowledge with machine learning methods to identify diagnostically relevant biomarkers that distinguish between polyneuropathy and myopathy. Robust evaluation metrics such as F1-macro and area under the precision-recall curve (AUC-PR) are used to address class imbalance and accurately assess model performance in our rare disease dataset. The study demonstrates that machine learning models can effectively differentiate between PN and MY patients using low-depth EHR data, potentially accelerating diagnostic processes for these rare neuromuscular diseases.

[30-P2-16] Lesion Network Mapping based Convolutional Neural Networks: Predictive Performance and Interpretability

Presenter: Matthias Becher

Authors: Matthias Becher [1]; Ana Sofía Ríos [2]; Yunyou Tang [2]; Ahmed Khalil [2]; Matthias Endres [3]; Heinrich Audebert [2]; Anna Kufner [2]; Ulrike Grittner [4]

Affiliations: [1] Charité - Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Institut für Biometrie und klinische Epidemiologie, Germany, Berlin; [2] Charité - Universitätsmedizin Berlin, Corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Klinik für Neurologie mit Experimenteller Neurologie Germany, Berlin; [3] Department of Neurology with Experimental Neurology, Charité - Univeristätsmedizin Berlin, Berlin; [4] Insitute of Biometry and Clinical Epidemiology, Charité Universtiaetsmedizin Berlin, Germany, Berlin

Introduction

Stroke can lead to a wide range of symptoms including acute motor impairment, post-stroke depression or cognitive impairment. Anticipating these outcomes early on and understanding their underlying causes would enable clinicians to initiate appropriate targeted treatments, which could not only help reduce the severity of symptoms but also improve long-term recovery. Convolutional neural networks (CNNs) are often used for these types of tasks and commonly rely on either computed tomography (CT) or magnetic resonance imaging (MRI) images as data for training. However, it's increasingly understood that complex neurological symptoms often arise from diaschisis, that is, from disconnections between brain regions, rather than being attributable to a single focal neuroanatomical substrate.

Methods

Lesion-network mapping (LNM) uses normative rs-MRI data to derive so-called connectivity maps indicating functional connections from brain lesions. In this project we investigated 1) whether CNNs trained on connectivity maps (LNM-CNN) outperform those trained on lesion masks (Lesion-CNN) in predicting post-stroke motor-impairment, and assessed 2) whether layer-wise relevance propagation (LRP) can be employed to extract valuable insights into model decisions. Finally, 3) we compared regional importance extracted from the LNM-CNN to those of a current standard procedure (permutation-based LNM).

Results

Both models performed similarly (area und receiver operating characteristic curve [AUC] both = 0.76), though the LNM-CNN showed higher recall and F1-Score (0.82 vs. 0.64, 0.75 vs. 0.69) indicating a higher sensitivity in identifying affected patients as well as better balance between precision and recall. LRP showed that importance often clustered near the lesion area, even in connectivity-based models, providing valuable context for the interpretation of the prediction results. The LNM-CNN revealed broader importance across cortical, subcortical, and white matter regions, while the permutation-based analysis predominantly identified cortical regions as important.

Discussion

Overall our study showed that in predicting acute motor impairments after a stroke, connectivity-based features slightly improved predictive performance compared to lesion masks in CNN models. Additionally, our results show that connectivity-based CNN models may offer complementary insights into the functional impact of brain lesions and highlight the need for further investigation of their clinical relevance.

[30-P2-17] From Detection to Report: Evaluating Open-Source Vision and Multimodal Models for Tuberculosis Chest X-ray Classification and Indonesian Report Generation

Presenter: Vanya Valindria

Authors: Vanya Valindria [1]; Giovan Christoffel Sihombing [2]

Affiliations: [1] Monash University, Jakarta; [2] Universitas Indonesia, Jakarta

Introduction

Tuberculosis (TB) remains a major public health challenge in Indonesia, where access to radiological expertise is limited. Open-source models such as X-Raydar and MedGemma are trained on large-scale datasets from high-income countries and primarily English-language reports. Their ability to generalize to Indonesian populations and generate clinically meaningful reports in Bahasa Indonesia remains underexplored.

Methods

We evaluated 500 chest X-rays from Indonesian patients annotated by radiology experts across 14 TB-related abnormalities. Classification performance was assessed using accuracy, precision, recall, and F1-score against expert ground truth for three models: MedGemma 27B, MedGemma 4B, and X-Raydar. Report generation quality in Bahasa Indonesia was evaluated separately by comparing generated reports to radiologist reports using BLEU, METEOR, and ROUGE-L. This was conducted for MedGemma (end-to-end) and a hybrid pipeline combining X-Raydar classification with MedGemma-based report generation.

Results

At the class level, X-Raydar achieved the highest accuracy in multiple abnormalities (e.g., cardiomegaly: 0.902 pleural effusion: 0.826), while MedGemma 27B consistently achieved the highest recall, frequently reaching 1.000 (e.g., cavity, consolidation). However, MedGemma 27B showed low precision (e.g., apical changes: 0.129), indicating over-detection. MedGemma 4B demonstrated lower overall performance, with reduced recall and F1-scores compared to 27B.

Macro-level analysis across all abnormalities shows that X-Raydar achieved the highest mean accuracy (0.60) and precision (0.42), while MedGemma 27B achieved the highest mean recall (0.85). Mean F1-scores were comparable between MedGemma 27B (0.44) and X-Raydar (0.43), reflecting different trade-offs between sensitivity and specificity.

For report generation, MedGemma outperformed the hybrid pipeline (BLEU: 0.605 vs 0.522 METEOR: 0.122 vs 0.104 ROUGE-L: 0.162 vs 0.145), indicating stronger cross-lingual generation capability.

Discussion

The results highlight complementary strengths between model types and scales. X-Raydar provides more balanced and clinically reliable abnormality detection, while MedGemma excels in generating coherent radiology reports in Bahasa Indonesia despite being trained on English-language data. However, discrepancies between higher BLEU and lower METEOR/ROUGE scores suggest that generated reports reproduce common phrases but lack deeper structural and semantic alignment with expert reports. Overall, open-source models show strong potential for TB screening and reporting in low-resource settings. However, effective deployment requires hybrid system design, local calibration, and careful balancing of precision and recall.

[30-P2-18] Federated Learning for Privacy-Preserving Chest X-Ray Classification in an Indonesian Clinical Setting: A Multi-Framework Evaluation

Presenter: Vanya Valindria

Authors: Vanya Valindria [1]

Affiliations: [1] Monash University, Jakarta

Introduction

Federated learning (FL) has emerged as a promising paradigm for developing collaborative medical artificial intelligence systems without requiring the sharing of sensitive patient data. This approach is particularly valuable in healthcare environments where privacy regulations are strict and data interoperability is limited. By enabling multiple institutions to jointly train models while keeping data locally stored, FL addresses both ethical and technical barriers to large-scale medical AI. In this study, we conduct a deployment-oriented evaluation of several state-of-the-art FL frameworks for a real-world medical imaging task: chest X-ray classification. Our goal is to assess their performance, efficiency, and practicality compared to traditional centralized learning approaches.

Methods

We evaluated four prominent FL frameworks: Fed-BioMed, MEDfl, OpenFL, and NVFlare. Each framework was used to train the X-Raydar-CV model on distributed local Indonesian chest X-ray datasets across multiple client sites, simulating real-world healthcare settings. A centralized training approach, where all data is pooled into a single location, was used as a baseline for comparison. Model performance was assessed using standard classification metrics, including ROC-AUC, PR-AUC, Micro-F1, and Macro-F1, with particular attention to handling imbalanced medical data. In addition, we measured computational efficiency by comparing total training time across all approaches.

Results

The results show that all federated learning frameworks significantly outperform the centralized baseline across all evaluation metrics. Improvements were especially notable in scenarios involving imbalanced datasets, where FL approaches demonstrated stronger generalization and robustness. On average, FL methods achieved nearly a twofold increase in key classification metrics such as PR-AUC and F1 scores. However, this performance gain came with increased computational cost, as federated training required approximately 3.7 times longer than centralized training due to communication overhead and distributed coordination.

Discussion

These findings highlight a favorable trade-off between computational efficiency and predictive performance when using federated learning in medical applications. Despite longer training times, the substantial improvements in model accuracy and robustness make FL a compelling solution for privacy-preserving healthcare AI. Importantly, the study demonstrates the feasibility of deploying FL frameworks in resource-constrained environments, including low- and middle-income countries, where centralized data collection may be impractical or prohibited.

[30-P2-19] A Knowledge Integration Framework (KIF) to support model-based simulated patients and virtual cohorts generation for rare and pediatric diseases

Presenter: Giulia Monchietto

Authors: Giulia Monchietto [1]; Sarah Zohar [1]; Karen Sinclair [2]

Affiliations: [1] Inserm, Université Paris Cité, Inria, HeKA, Paris; [2] Novartis Pharma AG, Basel

Introduction

Generation of virtual patients (VPs) is an attractive approach in rare and pediatric diseases settings, where patients do not exist at numbers required to conduct powered clinical trials. The objective of this work is to initiate a Knowledge Integration Framework (KIF) to reduce uncertainty in virtual patients generation. The framework links the available knowledge types (drug, disease and population knowledge) and knowledge sources (clinical trials, publications, non clinical data, experts elicitation) in the target or similar indication, to the appropriate existing methodologies that can be used to inform and qualify model, or model ensemble, chosen for VP generation.

Methods

VPs could be generated following a six-step workflow, in alignment with M15 guidelines: (1) Objective statement, (2) Assessment of knowledge available: type and sources of knowledge from target or similar indications, (3) Methodological selection: leveraging the scenario's knowledge existing applicable methods can be selected from the KIF, (4) Model qualification: assessing candidate models to evaluate robustness and associated prediction capability, (5-6) VP generation with the qualified model and uncertainty quantification will be the focus of future work, and will provide full measures of model risk, allowing regulators to understand the level of confidence in the evidence generated. The workflow is iterative, and if finally the model cannot be qualified for the intended use, additional data should be collected.

Results

We present two KIF applications, where the objective is to characterize variability in PK of psoriatic arthritis patients treated with Secukinumab, emulating a rare disease scenario, using a popPK model built on plaque psoriasis patients and limited trough concentration-time data are available from the target disease.

Discussion

Results show a qualified model with reduced uncertainty leveraging available sources of knowledge, for VP generation, in line with ICH M15 guidelines. While KIF is initiated with population pharmacokinetic (PopPK) models as the primary driver for VP generation, it is designed to be extendable to accomodate other modelling approaches, e.g. PKPD, PBPK, AI based models, and knowledge scenarios, e.g. real world data available, in future work.

[30-P2-20] Data-driven Identification of Socioeconomic Variables to Improve Prediction of Premature Mortality in Germany

Presenter: Mareile Beernink

Authors: Mareile Beernink [1]; Christopher Gundler [1]

Affiliations: [1] Institute for Applied Medical Informatics, University Medical Center Hamburg-Eppendorf, Hamburg

Introduction

Socioeconomic status (SES) is commonly quantified using expert-defined indices. In Germany, the German Index of Socioeconomic Deprivation (GISD) serves as an established standard, aggregating nine variables across income, education, and employment at district level. Studies confirm a strong association between SES and premature mortality, though model fit may be further extended using the extensive INKAR dataset. This study investigates which additional socioeconomic variables may improve the prediction of premature mortality.

Methods

We employed linear regression to compare six input configurations predicting premature mortality, evaluated by the coefficient of determination R^2: (1) the aggregated GISD score, (2) the nine GISD variables, (3) a filtered INKAR variable set of 132 variables excluding age-, sex-, population size-, and vital statistics-related variables, (4) the nine most predictive INKAR variables selected via Ridge regression (α=1.5) by coefficient magnitude, and (5, 6) the raw GISD variables extended by one and three INKAR variables. We sourced all data from the publicly available INKAR dataset from 2013-2021 at district level.

Results

Predictive performance ranged from median R^2=0.56 (GISD score IQR: 0.55-0.57) to 0.86 (INKAR set IQR: 0.84-0.86). Reducing the INKAR set to the nine most predictive variables yielded a median R^2 of 0.72 (IQR: 0.69-0.73). The GISD variables reached a median R^2 of 0.68 (IQR: 0.67-0.70). Adding the single most influential INKAR variable (Beschäftigte am WO mit Berufsabschluss) increased median R^2 by 7.35% to 0.73 (IQR: 0.73-0.75) combining it with Anteil Minijobs an den Beschäftigungsverhältnissen and Erwerbstätige im Tertiären Sektor raised median R^2 to 0.76 (IQR: 0.76-0.78), an 11.18% improvement. Likelihood-ratio tests confirmed statistically significant improvements in model fit for both additions ($p < 0.001$).

Discussion

The consistently high R^2 values reflect a strong relationship between SES and premature mortality. The aggregated GISD score, despite its interpretive simplicity, showed the weakest model fit. Extending the GISD variable improved predictive performance by up to 11.18%, a single variable alone accounting for 7.35%. These findings suggest that expanding the GISD variable structure could enhance predictive accuracy and support more robust analyses of socioeconomic health inequalities.

Future studies examining additional health endpoints are needed to validate and generalize the proposed variable selection.

[30-P2-21] Bayesian-Optimized Model Merging for Disease Risk Prediction in Small-Area Settings

Presenter: Shuzo Ishizaka

Authors: Shuzo Ishizaka [1]; Shunsuke Minusa [1]; Yasutaka Hasegawa [1]; Takanobu Osaki [1]; Hideyuki Ban [1]

Affiliations: [1] Hitachi, Ltd., Research Development Group, Tokyo

Introduction

In Japan, municipal insurers are implementing Data Health Plans, which aim to deliver health services more effectively and efficiently through analyses of healthcare data such as health checkup and medical claims data for health needs assessment and disease prediction. These analyses are often conducted at regional or subregional levels, but in small-area settings the target-area sample size may be too limited for stable estimation and accurate prediction. This study aimed to improve disease risk prediction in small-area settings by integrating statistical models learned from other regions into a target-area model.

Methods

We propose a model-merging framework that combines predictive models across regions by integrating their parameter distributions. Bayesian logistic regression was fitted independently in each region to obtain posterior distributions of regression coefficients. These posterior distributions were then merged using a Gaussian mixture model, with regional contributions weighted by mixture proportions. The mixture proportions were optimized by Bayesian optimization to maximize predictive performance in the target area. We evaluated the method using simulation data reflecting patterns reported in prior epidemiological studies.

Results

We compared predictive performance between a model trained solely on target-area data and a merged model that integrates information from a larger source region. In a diabetes onset prediction task using age, BMI, and HbA1c as predictors, under a data-scarce target-area condition (n=1,000), the target-only model achieved an AUC of 0.647, whereas the merged model achieved an AUC of 0.659. These results indicate that Bayesian-optimized parameter merging can improve predictive performance when data availability in the target area is limited.

Discussion

In small-area settings, limited sample sizes can make regression coefficients unstable and weaken predictive performance. By borrowing information from other regions through parameter-distribution merging, the proposed framework stabilizes estimation and improves discrimination. Model merging may therefore provide a practical strategy for disease risk prediction and support data-driven health services in small areas.

[30-P2-22] Advancing Small Area Estimation via Gradient Boosting: Evidence from Bangladesh

Presenter: Md Hasinur Rahaman Khan

Authors: Md Hasinur Rahaman Khan [1]

Affiliations: [1] Institute of Statistical Research and Training, University of Dhaka, Dhaka

Introduction

Reliable small area estimates (SAEs) are critical for evidence-based policymaking, particularly in low- and middle-income countries where administrative-level decision-making requires granular indicators. However, conventional survey-based estimators often yield imprecise estimates at disaggregated levels due to limited sample sizes and restrictive model assumptions. Classical SAE approaches, including Fay-Herriot and Elbers-Lanjouw-Lanjouw (ELL) models, partially mitigate these issues but remain constrained in capturing nonlinearities and leveraging high-dimensional auxiliary information.

Methods

This study develops a machine learning driven SAE framework based on gradient boosting algorithms, with a primary focus on LightGBM and comparative implementation of XGBoost. Using nationally representative survey data from Bangladesh, integrated with census, administrative, and geospatial auxiliary sources, we construct unit-level predictive models to generate domain-level indicators through aggregation. Survey design features are incorporated via probability-weighted training procedures. To address the fundamental challenge of uncertainty quantification in machine learning-based SAE, we implement a cluster bootstrap approach to estimate mean squared error (MSE) and construct confidence intervals.

Results

Empirical findings indicate that gradient boosting-based SAE methods substantially outperform classical benchmarks in predictive accuracy, calibration, and spatial granularity. Among the competing approaches, LightGBM consistently demonstrates superior performance across multiple development indicators, including health and education outcomes. The proposed framework yields stable and reliable district-level estimates, while the bootstrap-based uncertainty measures exhibit coherence with design-based inferential requirements.

Discussion

The integration of machine learning with SAE methodology offers a powerful and flexible alternative to traditional approaches, particularly in data-rich yet sample-constrained environments. This study provides both methodological advancement and practical guidance for incorporating machine learning-based SAE into official statistical systems, thereby enhancing the production of high-resolution, policy-relevant indicators in contexts such as Bangladesh.

[30-P2-23] Mapping disparities in air conditioning using geospatial prediction models in low and middle income countries

Presenter: Amber Vayani

Authors: Amber Vayani [1]; Tim Lucas [1]

Affiliations: [1] University of Leicester, Biostatistics Research Group, Division of Public Health and Epidemiology, School of Medical Sciences, Leicester

Introduction

Extreme heat is an increasing public health concern, yet access to air conditioning (AC) as a protective measure remains unevenly distributed worldwide. In low and middle income countries (LMICs), household data on AC ownership is limited, making it difficult to identify vulnerable populations at risk. This study addresses this gap by developing a predictive framework to estimate the spatial distribution of access to AC across LMICs in Africa and Asia.

Methods

Household level AC ownership data were obtained from Demographic and Health Surveys and linked to spatial covariates derived from global rasters. These included climatic and socioeconomic indicators. Predictive models were trained using machine learning approaches with repeated cross validations. This was then used to create a continental-scale prediction of estimated AC prevalence at a 5 km spatial resolution for the year 2023.

Results

Internal validation showed a high model performance with a MAE of 0.08 and an R2 of 0.73. However, under leave-country-out cross validation the model performed poorly with a MAE of 0.09 and R2 of 0.24 suggesting heterogeneity between countries. Spatial predictions indicate substantial geographic variation in AC access with very low predicted ownership across large parts of Sub-Saharan Africa. In contrast, higher predicted prevalence was concentrated in parts of South and Southeast Asia, especially in urban and economically developed areas.

Discussion

The results highlight how big the disparities are in adaptive capacity to extreme heat, with many high exposure regions lacking access to AC. The modelling approach used in this study provides a scalable method of estimating infrastructure gaps to areas with limited data. These findings can help support targeted interventions to protect health in a more equitable way. Future studies can include temporal trends to reflect AC ownership over time, improve uncertainty estimation and integrate additional data sources.

[30-P2-24] The Prediction Paradox in a No-Show Algorithm: to Show or not to Show?

Presenter: Heleen Brüggen

Authors: Heleen Brüggen [1]; Maarten van Smeden [1]; Wouter van Amsterdam [1]; Oisín Ryan [1]; Ewout Steyerberg [1]

Affiliations: [1] Utrecht University Medical Center, Utrecht

Introduction

The prediction paradox occurs when the prediction of an implemented algorithm alters the outcomes it aims to predict: the model improves patient outcomes, but discrimination decreases since high risk patients' risk decreases (Perdomo et al., 2020 Van Amsterdam et al., 2025). Our university medical center uses an algorithm to predict patients' no-show risk for out-patient clinic appointments, reminding high risk patients of their appointment and reducing the number of no-shows (Aij et al., 2024). This study examines the presence and extent of this prediction paradox within the implementation of UMC Utrecht's no-show algorithm, identifying how naive monitoring can lead to wrong conclusions, and evaluates reliable metrics for performance.

Methods

We used appointment data (2016-2026) from all out-patient clinics at UMC Utrecht including pre-deployment, internal RCT, and post-deployment data with regular retraining of the model, to investigate the prediction paradox in the no-show algorithm. We monitored discrimination, calibration, and predicted risk change for individual patients before and after retraining, as well as the predicted risk of no-show for patients who were and were not called.

Results

Post-deployment, discrimination and calibration of the algorithm declined before retraining, suggesting a decline in performance. Calibration deteriorated for called patients with higher predicted risk, where predicted risk exceeded observed risk. The performance remained stable when evaluating only the patients not called pre- and post-retraining.

Discussion

The discrimination and calibration differences pre- and post-deployment and between the patients who were called and weren't called demonstrate the prediction paradox. Furthermore, standard naive monitoring of the algorithm is unable to account for the prediction paradox. Reliable metrics are related to monitoring only non-called patients, where the prediction paradox is not induced.

Aij, K., Knoester, J., and Werkhoven, B. (2024). An artificial intelligence-based model to reduce the no-show rate in outpatient clinics of an academic hospital.

Perdomo, J., Zrnic, T., Mendler-Dünner, C., and Hardt, M. (2020). Performative Prediction. In Proceedings of the 37th International Conference on Machine Learning, pages 7599-7609. PMLR.

Van Amsterdam, W. A., Van Geloven, N., Krijthe, J. H., Ranganath, R., and Cinà, G. (2025). When accurate prediction models yield harmful self-fulfilling prophecies. Patterns, 6(4):101229.

P3 - Clinical Trials

Poster area: P3

[30-P3-01] From Change from Baseline to Disease Modeling: A Methodological Evaluation of Endpoint Metrics

Presenter: Cornelia Ursula Kunz

Authors: Cornelia Ursula Kunz [1]; Benjamin Lang [1]

Affiliations: [1] Boehringer Ingelheim, Biberach an der Riss

Introduction

Change-from-baseline endpoints are widely used in clinical studies to quantify progression or response. However, absolute change measures often depend on the baseline value, which can distort clinical interpretation and statistical inference. This baseline dependency affects patient ranking, introduces scale inconsistencies, and can violate key statistical assumptions. This work evaluates alternative endpoint formulations and demonstrates whether relative change offers more robust methodological properties across common progression models.

Methods

Analytical derivations and structured simulations were used to study three families of progression models frequently applied in longitudinal settings: linear, exponential, and rational models. For each model, we examined absolute, relative, and percentage change based on the original measurement and reciprocal transformations thereof. Metrics were evaluated according to (i) dependence on baseline values, (ii) interpretability, (iii) distributional behavior, (iv) compliance with independence and identical distribution (i.i.d.) assumptions, (v) behavior under baseline imbalance, and (vi) stability of patient ranking across models.

Results

Absolute change metrics showed strong baseline dependency: identical absolute changes implied different magnitudes of progression depending on the starting value, leading to contradictory rankings across models. Transforming measures using reciprocals did not resolve these issues absolute change of reciprocals remained bounded by the baseline, violating i.i.d. assumptions and inflating type I error under baseline imbalance.

In contrast, relative change - defined as the ratio of follow-up to baseline - was (mainly) baseline-independent and produced consistent rankings across linear, exponential, and rational models. Relative change aligned naturally with log-transformed endpoints and had direct mathematical links to integrative progression measures such as area-under-the-curve. Importantly, relative change is more likely to have no prognostic correlation with baseline values and allowed endpoint distributions

to remain identical across individuals, supporting valid hypothesis testing even under heterogeneity in baseline levels.

Discussion

Relative change provides a statistically principled alternative to absolute change for quantifying longitudinal progression. It minimizes baseline-driven bias, ensures fair and stable ranking, and maintains desirable properties across a wide range of progression models. For statistical methodology and applied research relying on change-from-baseline metrics, relative change - and log-based formulations - should be considered the preferred option over absolute change or inverted-scale differences.

[30-P3-02] Regression to the Mean and Repeated Measures in Randomized Clinical Trials

Presenter: Dominik Grathwohl

Authors: Dominik Grathwohl [1]

Affiliations: [1] Société des Produits Nestlé S.A., Nestlé Research, Clinical Research Unit, Lausanne 26

Introduction

Regression to the mean (RTM) is typically regarded as an issue in epidemiology. In randomized clinical trials, however, it is often assumed that RTM is irrelevant because inference is always made in comparison to a control group. Nevertheless, even in randomized clinical trials, RTM can occur and may lead to biased treatment effects if not properly addressed. In this investigation we focused on models which are widely used for repeated measures analysis such as repeated measures analysis of variance (RM-ANOVA) and different kind of mixed models. We provide a systematic comparison of the different models, and we illustrate and compare their performance.

Methods

We generated a dataset of two groups in parallel (n/group=80) with a baseline and two follow up measurements. We introduced a difference at baseline which gets smaller with time exactly by the correlation between visits, thus displaying only an RTM effect (RTM-dataset). We investigated: Full multivariate mixed model Mixed model on changes scores from baseline RM-ANOVA on change scores from baseline Mixed model correcting for baseline RM-ANOVA correcting for baseline Mixed model correcting for baseline and a baseline-visit interaction

Results

The false-positive rates for detecting a treatment effect in the RTM-dataset were: 61.7% - Full multivariate mixed model 44.4% - Mixed model on change scores from baseline 75.6% - RM-ANOVA on change scores from baseline 7.7% - Mixed model correcting for baseline 11.9% - RM-ANOVA correcting for baseline 6% - Mixed model correcting for baseline and a baseline-visit interaction

Discussion

Since the RTM-dataset contains only an RTM effect and no treatment effect, the correct model should have a false-positive rate of 5%. All investigated models fail to meet this criterium except the mixed model correcting for baseline and a baseline-visit interaction. Since this model has also a decent power in the presence of a treatment effect, it is our recommendation for repeated measures in randomized clinical trials.

[30-P3-03] Markov Ordinal State Transition Models for Time Out of Hospital and Quality of Life in Patients with Advanced Cancer

Presenter: Johannes Schwenke, Johannes Schwenke

Authors: Johannes Schwenke [1]; Perrine Janiaud [2, 3]; Lars G Hemkens [2, 3, 4]; Matthias Briel [1]; Benjamin Kasenda [5]; Giusi Moffa [5]

Affiliations: [1] Department of Clinical Research, CLEAR-Methods Centre, University Hospital Basel and University of Basel, Basel; [2] Department of Clinical Research, University of Bern, Bern; [3] Pragmatic Evidence Lab, Research Center for Clinical Neuroimmunology and Neuroscience (RC2NB), University Hospital Basel and University of Basel, Basel; [4] Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford; [5] Department of Medical Oncology, University and University Hospital Basel, Basel

Introduction

Markov ordinal state transition models (MOST) offer a framework for analyzing longitudinal patient states in clinical trials by modeling transitions between ordered health categories over time. We applied MOST in the LIQPLAT trial, a registry-based clinical trial that used a random invitation design to evaluate ctDNA-guided management versus usual care in patients with advanced solid cancer at the University Hospital Basel, Switzerland (N=298 NCT06367751). The trial exclusively uses routinely collected data, which presents both opportunities and challenges for MOST.

Methods

We specified two separate Bayesian MOST analyses: one for weekly hospitalization states (5 levels, including death as an absorbing state) and one for quality of life measured using the 7-point EORTC global health status item. Estimands were defined as differences in expected weeks spent in favorable states over 6 months between patients invited to ctDNA-guided management and patients receiving usual care, derived from model-based state occupancy probabilities. To inform model specifications in terms of Markov order and functional form over time, we used historical routinely collected data. Because the trial uses routinely collected data, we could investigate higher-order Markov structures, as data are also available prior to baseline. We used a second-order model with patient-level random effects for the hospitalization outcome. For quality of life, the sparsity of observations was particularly challenging: when non-absorbing states are incompletely observed but death is always known, the model systematically overestimates mortality. We addressed this by restricting the quality of life analysis to a principal stratum of survivors, as we expect little to no treatment effect on mortality.

Results

Follow-up completes in June 2026. We will present full comparative results on both outcomes, including agreement between model predictions based on historical data and trial observations.

Discussion

Historical data from the same data generation process were useful for pre-specifying MOST models. The differential observability of absorbing versus non-absorbing states in sparse longitudinal data is a methodological challenge that deserves further investigation and is relevant beyond our specific trial setting.

[30-P3-04] Title: Simulation-based sample size calculations for modelling heterogeneous treatment effects using the pmsims R package.

Presenter: Gordon Forbes

Authors: Gordon Forbes [1]; John Hodsoll [1]; Oyebayo Ridwan Olaniran [2]; Sarah Markham [2]; Diana Shamsutdinova [2]; Daniel Stahl [2]; Ewan Carr [2]

Affiliations: [1] King's College London, London; [2] King's College London, London

Introduction

The existence of heterogeneous treatment effects is of interest to patients, regulators, and scientists and central to personalised medicine. However, reliable detection and estimation requires adequate sample sizes. If sample sizes are too small, models may have inadequate power to detect heterogeneous treatment effects, may not reliably identify the variables that drive heterogeneity, and may perform poorly in predicting an individual's treatment effect due to overfitting.

Over the last decade, there have been advances in methods for assessing heterogeneous treatment effects, including those that facilitate the simultaneous analysis of multiple effect modifiers or avoid linearity assumptions. Model performance metrics have also been developed to assess how well an individual's treatment effect can be predicted, complementing approaches that estimate the effect of a particular variable on treatment effect.

Currently, however, there are no tools available for estimating the minimum sample size required for recently developed analysis approaches. In this presentation, we show how the pmsims package - a flexible, model-agnostic, simulation-based sample size tool - can be used to estimate the minimum sample size required for modelling of heterogeneous treatment effects and demonstrate the validity of the tool's output.

Methods

We compare minimum sample sizes across different analysis approaches, including penalised regression, model-based recursive partitioning, and causal forests. We estimate the sample size required under different assumptions of data complexity and for different model evaluation metrics, including the power of statistical tests to detect heterogeneous treatment effects, the ability to identify key

variables, and model performance in predicting individual treatment effects, using the calibration-for-benefit and c-statistic-for-benefit metrics. We assess the stability of output and whether model performance is at the expected level at the estimated minimum sample sizes.

Results

The results will highlight the sample size benchmarks across a range of models, evaluation metrics and data-generating assumptions. In addition, we will share the evaluation of the validity of the pmsims package for estimating minimum sample sizes for heterogenous treatment effects.

Discussion

This work addresses an important gap in sample size methodology, with implications for researchers analysing heterogeneous treatment effects and for clinical study design. The work highlights the flexibility of the pmsims pacakge (https://pmsims-package.github.io/pmsims/index.html) for sample size estimation.

[30-P3-05] Improving design and reporting for biomarker-negative populations in phase III marker-stratified trials

Presenter: Mana Watanabe

Authors: Mana Watanabe [1]; Shigeyuki Matsui [2]

Affiliations: [1] Department of Clinical Biostatistics, Kyoto University School of Public Health, Kyoto; [2] Department of Biostatistics and Data Science, Kyoto University School of Public Health, Kyoto

Introduction

Biomarker-stratified phase□ oncology trials are usually designed to demonstrate efficacy in the overall population and/or biomarker-positive subgroup. Even when biomarker-negative and intermediate subgroup results are reported, they are often not adequately evaluated or presented in an interpretable manner. Moreover, no established criteria exist for evaluating biomarker-defined subgroups or for designing trials that ensure sufficient information for these populations. The objective of this study was to propose a statistical framework and practical guidelines for the design and reporting of biomarker-stratified phase□ trials that enable transparent estimation across clinically relevant subgroups.

Methods

We developed a framework that separates hypothesis testing from estimation and introduces complementary biomarker subgroup definitions. The framework was examined in representative scenarios motivated by phase□ immune checkpoint inhibitor trials in non-small cell lung cancer, assuming three biomarker-defined subgroups (low, intermediate, and high expression) and event-driven analysis for time-to-event endpoints. At the design stage, subgroup interpretability was evaluated by expected precision, quantified by the confidence interval width for log(HR), and by the operating characteristics based on the posterior probability of clinically meaningful treatment effects Pr(HR < HRc

data) under no treatment effect (HR=1), where HRc is a prespecified clinically meaningful threshold (e.g.,0.8).

Results

Under typical phase trial sizes, precise estimation within narrowly defined subgroups(e.g., low-expression populations) was limited, with wider confidence intervals and only moderate assurance for ruling out clinically meaningful benefit. In contrast, broader complementary subgroup definitions, such as low plus intermediate expression, provided greater precision and higher assurance under the same overall sample size. Complementary binary partitions-low versus intermediate/high and low/intermediate versus high-offered distinct and mutually informative perspectives, reducing both overgeneralization and under-recognition of treatment effects.

Discussion

The criteria based on the estimation precision and the posterior probability of clinically meaningful effects should be considered in future design, conduct, and reporting of biomarker-stratified trials. Adoption of such guidelines may reduce both over- and under-interpretation of subgroup effects and support more evidence base use of biomarker driven therapies.

[30-P3-06] The RobinCar Family: R Tools for Robust Covariate Adjustment in Randomized Clinical Trials

Presenter: Dominic Magirr

Authors: Daniel Sabanés Bové [1]; Marlena Bannick [2]; Liming Li [3]; Mark Baillie [4]; Dominic Magirr [4]

Affiliations: [1] inferential.biostatistics GmbH / RCONIS, Basel; [2] Seattle Children's Therapeutics, Seattle; [3] AstraZeneca BioPharmaceuticals, Shanghai; [4] Novartis Pharma AG, Basel

Introduction

Covariate adjustment is a powerful statistical technique that can increase efficiency in clinical trials. Recent guidance from the FDA provided recommendations and best practices for using covariate adjustment. However, there has existed a gap between the extensive statistical literature on covariate adjustment and software that is easy to use and abides by these best practices.

Methods

We have developed the RobinCar Family, which comprises RobinCar and RobinCar2 (Bannick et al, 2026). RobinCar2 provides clinical trial practitioners with a lite version of the more research oriented RobinCar package with the most essential and well validated statistical methods, following good practices for engineering high quality statistical software packages. These two R packages enable covariate-adjusted analyses for continuous, discrete, and time-to-event outcomes that follow best practices. For continuous and discrete outcomes, the functions in the RobinCar Family facilitate traditional forms of covariate adjustment such as ANCOVA as well as more recent approaches like ANHECOVA, G-computation with generalized linear models and machine learning models, and

adjustment for a super-covariate (see Holzhauer, Adewuyi, 2023). Functions for time-to-event outcomes implement the covariate-adjusted log-rank test, the stratified covariate-adjusted log-rank test, and the marginal covariate-adjusted hazard ratio.

Results

We provide an accessible overview of the covariate-adjusted statistical methods in RobinCar and RobinCar2, and show a 7% increase in effective sample size compared to a baseline method in 22 post-hoc analyzed oncology trials.

Discussion

The RobinCar Family provides a comprehensive suite of robust statistical methods for covariate adjustment in randomized clinical trials, in line with recent FDA guidance (U.S. Food and Drug Administration, 2024). The packages RobinCar and RobinCar2 fill an important gap, facilitating the practical implementation of powerful new statistical methods. Finally, RobinCar2 is an open-source project of the American Statistical Association's Biopharmaceutical Section Covariate Adjustment Scientific Working Group. As such, it is supported by a community of developers across industry and academia, ensuring its long-term maintenance and sustainability.

Bannick M et al. The RobinCar Family: R Tools for Robust Covariate Adjustment in Randomized Clinical Trials. doi:10.48550/arXiv.2601.14498.

Holzhauer B, Adewuyi ET. " Super-covariates": Using predicted control group outcome as a covariate in randomized clinical trials. Pharmaceutical Statistics. 2023 22(6): 1062-1075. doi:10.1002/pst.2329

[30-P3-07] Advancing AI/ML-Based Prognostic Covariate Adjustment: Systematic Approaches to Multi-Dataset Integration for Enhanced Trial Efficiency

Presenter: Jan Wiemer

Authors: Jan Wiemer [1]; Clemence Rigaux [2]; Vivek Chowdhury [3]; Karl Köchert [1]

Affiliations: [1] Sanofi-Aventis Deutschland GmbH, Berlin; [2] Sanofi aventis Recherche et Developpement, Gentilly; [3] Sanofi Healthcare India Private Limited, Hyderabad

Introduction

AI/ML-based prognostic covariate adjustment is a promising approach to enhance clinical trial efficiency by including the prognostic score as covariate in efficacy analysis. The approach is endorsed by regulatory agencies (FDA 2023, EMA 2015) and demonstrated effective in retrospective analyses of e.g., Lecanemab trials. Despite regulatory acceptance, this methodology remains underutilized in practice. Its key advantage lies in capturing biological complexity - non-linear associations, non-monotonic patterns, and higher-order interactions - features missed by classical adjustment. However, extracting these dependencies reliably requires substantial sample sizes, a critical practical constraint.

Methods

Alzheimer's disease (AD) trials present both compelling need for efficiency gains - driven by long durations, expensive imaging, heterogeneous populations, and high endpoint variability - and unique methodological challenges. Chief among these is data availability: no single historical dataset may contain all relevant biomarkers or sufficient sample size for robust AI/ML model development.

We systematically explore strategies to address this using real patient data from TransCelerate member AD historical clinical studies, evaluating four complementary approaches: (1) developing separate prognostic scores from distinct historical datasets (2) harmonizing multiple studies via commonly available covariates to increase effective sample size (3) combining multiple AI/ML scores into unified 'super prognostic scores' and (4) comparing single comprehensive versus integrated multi-score approaches. AI/ML models are applied to TransCelerate member placebo arm data, with performance evaluated via Effective Sample Size Increase (ESSI) and cross-validated predictive accuracy.

Results

AI/ML models trained on individual TransCelerate member datasets achieved meaningful prognostic performance consistent with efficiency gains reported in the literature. Score integration strategies showed varying degrees of efficiency gain depending on data availability and biomarker overlap. Preliminary findings support the feasibility of the proposed framework in real AD trial data. Results will be illustrated through case examples comparing single-dataset versus multi-dataset approaches, with discussion of trade-offs and implementation considerations.

Discussion

Systematic multi-dataset integration and score combination strategies can unlock the potential of AI/ML-based prognostic covariate adjustment in complex therapeutic areas like AD, where comprehensive biomarker data is fragmented across studies. We provide practical guidance for implementing these methods in real-clinical trial settings, directly addressing a key barrier to broader adoption.

[30-P3-08] Enhancing Statistical Efficiency in ALS Trials Through Prognostic Covariate-Adjusted ANCOVA: A Simulation Study

Presenter: Sofia Azevedo, Paolo Messina

Authors: Sofia Azevedo [1]; Niccolò Totis [2]; Simona Catozzi [1]; Maud Beneton [1]; Mario Torchia [1]; Juliette Le Douce [3]; Sébastien Marie [3]; Annette Janus [3]; Yann Godfrin [3]; Daniel Roeshammar [1]; Pauline Bambury [1]; Paolo Messina [1]

Affiliations: [1] InSilicoTrials Technologies Spa, Trieste; [2] InSilicoTrials Technologies B.V., Amsterdam; [3] Axoltis Pharma, Lyon

Introduction

Randomized controlled trials in amyotrophic lateral sclerosis (ALS) are challenged by substantial inter-patient heterogeneity, limited sample sizes, and ethical concerns over placebo exposure. Serum neurofilament light chain (NfL) is an emerging biomarker for treatment response, but the potential for prognostic adjustment to improve statistical efficiency in short-term biomarker trials remains unclear. This study evaluates the impact of prognostic covariate-adjusted ANCOVA on trial power compared with unadjusted analyses.

Methods

We simulated a 3-arm trial reflecting the design of NCT06365216, with participants allocated in a 30:30:20 ratio to two dose levels of NX210c or placebo over a 6-week period. Serum NfL change from baseline (primary endpoint) was modeled as lognormally distributed, and a latent prognostic covariate (PS) was generated as the predicted primary endpoint derived from a semi-mechanistic model under a no-treatment assumption with correlations to the observed outcome of ρ = 0.1, 0.3, 0.5, and 0.7. Active treatment was modeled to reduce log NfL by -0.42 relative to placebo (in line with the planned study assumptions). For 1000 simulations per correlation scenario, we compared unadjusted and prognostic-adjusted ANCOVA, estimating change-from-baseline NfL, residual variance, confidence intervals, effect sizes, and statistical power for the active vs placebo comparison.

Results

Unadjusted analyses achieved median power ~0.83 across scenarios. Prognostic covariate-adjusted ANCOVA yielded progressively greater power gains with higher prognostic correlations: median power was ~0.83 (ρ = 0.3), 0.88 (ρ = 0.5), and 0.96 (ρ = 0.7). At low correlation (ρ = 0.1), adjusted power was slightly lower (0.80), consistent with limited additional prognostic information. Across scenarios, prognostic covariate-adjusted ANCOVA reduced residual variance, produced narrower confidence intervals, stabilized effect size estimates, and improved precision of treatment effect estimation, confirming that adjustment improves statistical efficiency predominantly when the covariate is strongly prognostic.

Discussion

For ALS trials using serum NfL as a primary endpoint, ANCOVA prognostic adjustment can meaningfully enhance statistical efficiency over unadjusted analyses when the prognostic covariate is highly correlated with the outcome. However, in short-term trials with moderate or weak biomarker correlations, power gains are more modest and limited. These results highlight prognostic covariate-adjusted ANCOVA as an effective and practical approach for optimizing statistical efficiency in biomarker-driven ALS studies.

[30-P3-09] Handling of Incomplete Baseline Covariates in Cluster-Randomised Trials: A Simulation Study

Presenter: Baptiste Leurent

Authors: Baptiste Leurent [1]; Clemence Leyrat [2]; Jennifer A Thompson [2]; Richard Hooper [3]; Helen Weiss [2]; Elizabeth Allen [2]

Affiliations: [1] University College London, London; [2] London School of Hygiene & Tropical Medicine, London; [3] Queen Mary University of London, London

Introduction

Adjustment for baseline covariates is recommended in the analysis of cluster-randomised controlled trials (cRCT) to increase power and mitigate chance imbalance. However, baseline are not always completely observed. There is currently a lack of guidance on how to best adjust for baseline variables when they are incomplete. The aim of this work was to compare the performance of different approaches to handle incomplete baseline covariate in cRCT, using simulations.

Methods

We simulated cRCTs with a continuous outcome and a partially observed baseline covariate, varying the number of clusters, clusters size, intra-cluster correlation coefficient (ICC), cluster-level and subject-level correlations (CC, SC) and proportion of missingness. The estimand was the mean difference between trial arms. We compared nine strategies when handling the incomplete baseline covariate: unadjusted, complete case analysis, adjusting for the baseline cluster means, mean-imputation (replacing the missing baseline values by the cluster means, or by the overall trial mean), and repeated-measures modelling. All methods were applied with or without adjustment for the baseline cluster-level mean. The bias, model-based standard error, confidence-interval coverage, and empirical efficiency were compared between methods. We applied the methods in a cRCT of an educational intervention in schools.

Results

All methods were unbiased and with correct coverage. Unadjusted or complete-case analyses were generally the least efficient. Repeated-measures model was generally the most efficient when both CC and SC were above 0.5. In other scenarios, repeated-measures model, adjusting for the cluster means or doing mean-imputation (either using the cluster or the overall mean) were generally similarly efficient. In the case study the precision was increased by 20 to 29% compared to an unadjusted analysis by adjusting for the cluster means, mean imputation, or using a repeated-measure model.

Discussion

Repeated-measures model was found to be the most efficient approach to adjust for incomplete baseline covariates when the correlation is high, but it needs to be specified appropriately. Simpler methods such as adjusting for the cluster means or mean imputation performed generally well and may be attractive in practice.

[30-P3-10] Clinical studies with treatment effect interactions in the main analysis: practical implications

Presenter: Erik Hermansson

Authors: David Svensson [1]; Magnus Kjaer [2]; Per Nyström [2]

Affiliations: [1] Statistical Innovation, AstraZeneca R&D, Gothenburg, Sweden, Gothenburg; [2] CVRM Biometrics, BioPharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden, Gothenburg

Introduction

In type 2 diabetes trials with HbA1c as a clinical outcome, there is evidence suggesting treatment effect heterogeneity with baseline HbA1c. When such effect modification is anticipated and clinically meaningful, including treatment-by-covariate interactions in the primary analysis can improve transparency and interpretability for patients, regulators, and other decision-makers. However, once treatment effects are allowed to vary across patient characteristics, both the framing of the scientific questions addressed by the trial and the interpretation of hypothesis tests become less straightforward.

Methods

We present an approach to hypothesis testing when treatment-by-covariate interactions are included in the primary analysis and highlight common misunderstandings, including confusion between evidence for an overall treatment benefit and evidence that treatment effects differ across patient subgroups.

Results

We demonstrate implications for the design and analysis plan of clinical trials, particularly potential artefactual results if using classical testing paradigms despite knowledge of treatment effect heterogeneity. Numerical examples are presented to illustrate key advantages and limitations of different approaches.

Discussion

We discuss implications and options for subgroup assessments when treatment effect heterogeneity is present, with a focus on clear pre-specification and coherent interpretation across subgroups. Furthermore, we address non-inferiority testing in this context, motivated by practical regulatory and clinical considerations, and provide practical recommendations for trial design and analysis.

[30-P3-11] Estimating average treatment effect on the treated in single-arm phase II clinical trials with external controls for time-to-event outcome

Presenter: Tianyuan Gu

Authors: Zhaojin Chen [1]; Jialiang Li [2]; Ross Andrew Soo [3, 4]; Bee Choo Tai [1, 5]; Tianyuan Gu [6]

Affiliations: [1] Saw Swee Hock School of Public Health, National University of Singapore, Singapore, Singapore; [2] Department of Statistics and Data Science, National University of Singapore, Singapore, Singapore; [3] Department of Haematology-Oncology, National University Cancer Institute Singapore, Singapore, Singapore; [4] Cancer Science Institute of Singapore, National University of Singapore, Singapore, Singapore; [5] Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore; [6] National University of Singapore, Saw Swee Hock School of Public Health, Singapore

Introduction

In oncology, single-arm phase II trials are commonly used to evaluate anticancer activity of therapeutic agents by comparing outcome to pre-defined threshold, often resulting in biased estimation of treatment effect and inflated type I error. This study aims to improve the current design by incorporating an external control arm of patients receiving standard of care. We evaluate the performance of causal inference methods when estimating the average treatment effect on the treated (ATT) for time-to-event outcome.

Methods

Inverse probability of treatment weighting (IPTW), one-to-one greedy nearest neighbour matching without replacement (NNM), NNM with caliper (NNM-C) and optimal full matching (OFM), were used to balance the baseline confounders. Cox proportional hazard regression with robust variance estimate was subsequently used to estimate the ATT. In the simulation study, time-to-event outcome were generated from Gompertz proportional hazard regression model, where three baseline variables followed independent Bernoulli distributions and censoring time followed a uniform distribution. Simulation study was conducted by varying the extent of imbalance, censoring percentage and treatment effect for sample sizes of 100 and 200. The results were compared with an additional setting mimicking a randomised trial with equal allocation. The methods were further illustrated using data from the PLASMA and AURA 3 trials.

Results

Comparable to the setting of a randomised trial, NNM-C had negligible bias, small mean square error and achieved 95% coverage probability for n =200. In the case of IPTW, these were attained for small to moderate imbalance. All methods except for NNM maintained one-sided type I error below 0.1. Although, NNM-C and IPTW demonstrated lower power than the randomised trial setting, both could reject the null hypothesis with at least 80% power for large treatment effect assuming small to moderate imbalance. Similar results were observed for NNM-C and IPTW when there was small imbalance for n = 100.

Discussion

NNM-C and IPTW are recommended for the estimation of ATT, for small to moderate imbalance and n =200. Both methods performed reasonably for small imbalance and n = 100. Larger sample size or number of available controls may be required to detect moderate treatment effect with adequate power.

[30-P3-12] Should we use post-intercurrent event data when estimating hypothetical estimands? A simulation study

Presenter: Remo Schmutz

Authors: Remo Schmutz [1, 2]; Frédérique Chammartin-Basnet [1, 2]

Affiliations: [1] Division of Clinical Epidemiology, Department of Clinical Research, University Hospital Basel, Basel; [2] University of Basel, Basel

Introduction

Under the ICH E9(R1) estimands framework, the hypothetical strategy targets the treatment effect in a world where the intercurrent event (IE) did not occur. Conventionally, post-IE observations are discarded, but recent work has shown that incorporating them via G-estimation can improve efficiency (Lasch et al. 2023). This efficiency gain, however, relies on an assumption that the IE affects all patients homogeneously. When this assumption is violated, the G-estimator becomes biased and methods discarding post-IE data remain unbiased. Wu et al. (2024) restored unbiasedness by extending the G-estimator to account for treatment-IE interactions. Whether any efficiency advantage remains after this correction is an open question. Bartlett and Daniel (2025) argue analytically that gains from using post IE data in this way will be modest in most scenarios. The present simulation study is the first to examine this question empirically.

Methods

We conduct a preregistered simulation study (preregistration in progress https://github.com/schmutzre/simstud IE) extending the data-generating mechanism of Bartlett and Daniel (2025) to include treatment-IE and covariate-IE interactions of varying magnitude, crossed with IE effect sizes, average IE rates, and rate differences between arms. Four estimators are compared: a pre-IE G-formula estimator (primary comparator) the Loh G-estimator (Loh et al. 2020) the interaction-corrected G-estimator of Wu et al. (2024) and an IPW estimator (Parra et al. 2025).

Results

Across all conditions we report bias, variance, mean squared error (MSE), and relative efficiency (RE) with Monte Carlo uncertainty. The primary focus is how RE of the interaction-corrected G-estimator varies with interaction magnitude, IE effect size, and IE rate, and whether any conditions yield RE meaningfully above 1. The results and the corresponding paper are expected to be finished at the time of the conference.

Discussion

This study provides the first empirical quantification of when, if ever, post-IE G-estimation retains a meaningful efficiency advantage after interaction correction, extending the analytical arguments of Bartlett and Daniel (2025) to a broad factorial grid. Results will offer practical guidance for analysts under the ICH E9(R1) framework.

Lasch et al. 10.1080/19466315.2022.2055633 Wu et al. 10.1002/sim.10212 Bartlett, J. W., & Daniel, R. M. arXiv:2504.00929 Loh et al. 10.1080/00273171.2019.1681251 Parra et al. 10.1080/19466315.2022.2081599

[30-P3-13] Beyond Toxicity: Statistical Applications to Adverse Event Data in Precision Oncology

Presenter: Dung-Tsa Chen

Authors: Dung-Tsa Chen [1]

Affiliations: [1] Moffitt Cancer Center, Tampa, Florida, Tampa

Introduction

This study introduces a bias-aware, high-dimensional analytical framework that fundamentally redefines the way adverse events (AEs) are analyzed in cancer clinical research. Traditionally, AEs are summarized in safety reports, often focusing on the worst grade experienced by a patient or reporting only single occurrences. However, this new framework elevates early-onset AEs to the status of quantitative biomarkers of treatment efficacy, recognizing their potential to predict therapeutic response.

Methods

Instead of simplifying AE analysis to the most severe event or a single episode, this framework integrates several AE parameters - including grade (low vs high), treatment attribution (differentiating between treatment-related and non-treatment-related AEs), and timing of onset (distinguishing early versus late occurrence relative to treatment initiation). It systematically defines and adjusts for dependencies among varying AE types via correlation analysis, followed by subgroup analysis. Additionally, the model incorporates survival landmark analysis at early time points for early-onset AEs to mitigate immortal time bias. A case study involving lung cancer is presented to illustrate the methodology.

Results

Statistical analysis reveals both intra- and inter-patient heterogeneity in AE profiles, with each patient experiencing unique patterns. AEs vary by grade, treatment attribution, and timing. The analytical model consolidates complex AE data into interpretable metrics for downstream analysis. Survival landmark analysis at one-month post-treatment shows that high burden of low-grade, treatment-related AEs is linked to improved survival, while non-treatment-related AEs predict worse outcomes. Due to observed negative correlation between low-grade and high-grade AEs, subgroup analyses stratified by non-treatment-related AEs confirm the positive survival association with low-grade, treatment-related AEs.

Discussion

This innovative framework modernizes AE analysis by transforming routine toxicity data into timely, reliable, and clinically actionable statistical biomarkers. It enhances efficiency, especially in immunotherapy and combination trials, supplying clinicians and researchers with robust tools to assess efficacy and guide patient care decisions.

[30-P3-14] Comparison of Multiple Imputation Approaches for Skewed Outcomes in Randomised Trials: A Simulation Study

Presenter: Jingya Zhao

Authors: Jingya Zhao [1]; Baptiste Leurent [1]; Gareth Ambler [1]

Affiliations: [1] University College London, London

Introduction

Missing outcome data are common in trials and can reduce efficiency and introduce bias. Multiple imputation (MI) is commonly used to handling missing data. However, standard MI methods for continuous variables typically assume normality, which is often unrealistic in practice, outcomes such as healthcare costs or alcohol consumption are often highly skewed. Alternative MI approaches, including predictive mean matching (PMM) and transformation-based methods, have been proposed for skewed variables. Using simulation, we compared methods for imputing missing skewed outcomes in randomised trials.

Methods

We simulated a two-arm trial setting with a single outcome variable with missing values and considered two estimands: the overall mean and the mean difference between treatment groups. Outcomes were generated from lognormal and gamma distributions with different levels of skewness. Missingness was imposed under missing completely at random (MCAR) and missing at random (MAR) with probability dependent on treatment arm. We compared seven methods: complete case analysis (CCA), linear regression imputation (Reg) and PMM, their log-transformed variants (LogReg and LogPMM), a non-parametric probability integral transform (PIT) approach, and a new bootstrap PMM method (BS-PMM). Performance was assessed in term of bias, relative error in standard error and 95% CI coverage. The methods were also applied to a clinical trial assessing healthcare cost.

Results

LogReg produced biased estimates even with lognormal outcomes, with greater bias under gamma outcomes. PMM and LogPMM produced approximately unbiased estimates but underestimated standard errors. In contrast, Reg and PIT performed well under both MCAR and MAR, PIT is particularly attractive for preserving skewness of the observed data.In a healthcare costs trial, LogReg severely overestimated the point estimate and its standard error. Other methods gave similar results, with PMM returned smaller standard errors than other methods

Discussion

PMM can under-estimate standard errors. Log-transformation for imputation is not recommended, as it can induce bias even for lognormal outcomes. In the settings considered, regression and PIT imputation performed well under MCAR and treatment-dependent MAR. While further work is needed to understand behaviour in more complex settings, this work highlights that imputation methods that appear attractive to handle non-normality should be used with caution.

[30-P3-15] RECIST oncology measurement error: an alternative sensitivity analysis to assess potential within-BICR measurement error

Presenter: Alexandra Hunt

Authors: Alexandra Hunt [1]

Affiliations: [1] Phastar, Sheffield

Introduction

FDA guidance discusses various endpoints to be considered in oncology clinical trials, with overall survival (OS) often considered the gold standard. An alternative endpoint to OS is progression-free survival (PFS), which is often considered as the primary endpoint. PFS is commonly analysed by adopting the blinded independent central review (BICR) process, whereby more than one reviewer will evaluate each subject. In the case of discordance between radiologists, an additional reader will adjudicate to decide upon an accepted record to be used in the final analysis. The FDA acknowledges that the subjective nature of PFS does have disadvantages such as bias, and therefore sensitivity analyses are encouraged to evaluate robustness of the results.

A typical sensitivity analysis used in oncology studies focuses on assessment bias, to compare the BICR results with local investigator findings. Whilst most studies will look to assess discordance between investigator and BICR assessments, less attention is given to discordance between BICR radiologist measurements. Beaumont and Iannessi (2023) identified a 21% progressive disease declared discrepancy rate from double read BICR lung cancer trials. Therefore, this poster will look to explore the conservative scenario of always selecting the progressive disease outcome in the case of BICR disagreements between radiologists.

Methods

This research utilises published PFS analysis results from a previous oncology clinical trial, to simulate a two-arm clinical trial. The simulation incorporates a visit schedule and accounts for progressive disease discrepancies between two readers. Hazard ratios are presented and median PFS calculated using the Kaplan-Meier technique.

Results

The poster will show the following:

Scenario one: results of the simulated trial based on published trial results Scenario two: results of the sensitivity analysis to assess the robustness of the scenario one results, whereby in the case of radiologist discordance, the progressive disease record is used in analysis.

Discussion

The research will show how scenario two, albeit more conservative, can form the basis of a robust sensitivity analysis in the context of BICR PFS analysis.

Beaumont, H. and Iannessi, A., 2023. Can we predict discordant RECIST 1.1 evaluations in double read clinical trials?. Frontiers in Oncology, 13, p.1239570.

[30-P3-16] Hidden in the Code: Implementation Variability and Its Impact on Clinical Trial Conclusions

Presenter: Nicole Close, Yaroslav Sud

Authors: Nicole Close [1]; Yaroslav Sud [1]

Affiliations: [1] EmpiriStat, Inc., International

Introduction

Statistical analyses in clinical trials are typically assumed to follow directly from protocol- and SAP-specified methods. In practice, however, many implementation decisions are resolved during programming, including visit window allocation, baseline derivation, exposure definitions, censoring rules, and treatment of incomplete data. While often regarded as operational details, these choices embed implicit assumptions that may alter the effective estimand and influence inferential conclusions. We investigate implementation variability as a systematic and under-recognised source of analytic divergence in clinical trials.

Methods

We evaluated recurring analysis situations in which multiple programming implementations remain fully compatible with identical SAP wording. Using examples derived from contemporary industry trials across therapeutic areas, we constructed parallel derivation strategies for visit assignment, baseline definition, and analysis population construction. Each implementation pathway was analysed using identical statistical models to isolate the impact of derivation logic on treatment estimates, uncertainty, and interpretability. Particular attention was given to identifying categories of implementation decisions most likely to influence inferential results.

Results

Across scenarios, alternative implementations produced consistent directional findings but materially different effect sizes, precision estimates, and subgroup interpretations. The greatest divergence arose when implementation rules affected subject inclusion, timing of measurements, or handling of incomplete data. These differences were sufficient in some cases to alter secondary conclusions or the perceived robustness of treatment effects, despite complete adherence to the written SAP. The findings demonstrate that implementation variability can act as a hidden layer of analytical flexibility within otherwise well-controlled analyses.

Discussion

Our results suggest that reproducibility and interpretability of clinical trial findings depend not only on statistical models but also on transparent specification of derivation logic. In alignment with the estimand framework articulated in ICH E9(R1), implementation rules should be recognised as part of the inferential structure of an analysis rather than purely technical execution steps. Explicit identification of high-impact implementation decisions offers a practical route to improving analytical transparency, reducing avoidable variability, and strengthening confidence in trial conclusions.

[30-P3-17] Variance of analysis - R package for clinical trials

Presenter: Reinhard Vonthein

Authors: Reinhard Vonthein [1]

Affiliations: [1] Universität zu Lübeck, Lübeck

Introduction

Statistical analysis relies on more or less or different, but similar assumptions. Such researcher degrees of freedom exist and may alter conclusions.[1, 2, 3] As a consequence, the ICH-E9 Addendum recommends sensitivity estimands that rely on slightly different assumptions.[4] As many trials in medicine, psychology and marketing ("AB-testing") are two randomized parallel groups with baseline and endpoint measurements, the code needed is supplied here.

Methods

The code used in the RCT NOVICE [5] was cautiously generalized. It takes data in wide format. One function produces the long and longer formats required by some other R packages. Other functions estimate the difference of location change between groups for dichotomous, ordinal, and measured variables assuming different models and distributions, i.e. from median test over ANCOVA to mixed generalized linear regression. The numerous estimands are presented in tables with correct wording together with a paragraph for the methods section, if used in a markdown file. Another function describes rater agreement.

Results

The R package varoan may be downloaded from github.[6] While installing it, the R packages needed for the calculations are installed, too. Different file formats are supported for import. A vignette contains the full analysis set of RCT NOVICE.[7] The output language may be switched between German and English. Forrest plots are produced for point and interval estimates that share a scale.

Discussion

While the choice of estimands may be abused for p hacking, it may serve the assessment of robustness of the prespecified primary estimand. Users of varoan may not reproduce all of the results of the studies with the many teams. No option for imputation has been integrated for now.

References

[1] Breznau et al. 2022 PNAS https://doi.org/10.1073/pnas.2203150119

[2] Silberzahn et al. 2018 Adv Meth Pract Psychol Sci https://doi.org/10.1177/2515245917747646

[3] Silberzahn&Uhlmann 2015 Nature https://doi.org/10.1038/526189a

[4] ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials - Step 5. EMA/CHMP/ICH/436221/2017 https://www.ema.europa.eu/en/ich-e9-statistical-principles-clinical-trials-scientific-guideline [21.01.2026]

[5] Thomaschewski et al. 2025 Int J Surg. DOI:10.1097/JS9.0000000000002304

[6] github.com/vonthein

[7] NOVICE Raw data. https://doi.org/10.5281/zenodo.10428830

P4 - Causal Methods and Epidemiology

Poster area: P4

[30-P4-01] Risk Evaluation for Neonates Born to Overweight and Obese Women in a Bavarian Maternity Hospital (Care Level IV)

Presenter: Christine Gall

Authors: Christine Gall [1]; Ulf Dammer [2]; Christian R Löhberg [2]; Sven Kehl [3]

Affiliations: [1] University of Erlangen-Nuremberg, Erlangen; [2] Department of Obstetrics and Gynaecology, St. Theresien Hospital Nürnberg, Nuremberg; [3] Klinik und Poliklinik für Frauenheilkunde und Geburtshilfe, Campus Großhadern und Innenstadt, München

Introduction

There are many reasons why pregnant women may prefer to give birth in smaller maternity units (Care Level IV facilities), including a calmer setting, more personalized care, and a less "clinical" atmosphere. As specialized obstetric and neonatal services are not available on-site, accurate information on an individual's risk profile is essential to support women in making a well-informed decision.

For women with obesity, delivery in a perinatal center (care levels I-II) is often recommended. However, obesity often co-occurs with other correlated factors, and unknown confounders might influence the choice of maternity hospital. To reduce referral-driven selection and confounding by care level, we use quality-assured routine data from a Bavarian care level IV hospital to assess the effect of a higher body mass index on fetal outcomes.

Methods

This retrospective cohort study included 5,616 women who delivered between 2016 and 2023 at a Bavarian care level IV maternity hospital. The primary outcome was neonatal transfer to a neonatal intensive care unit (NICU). Secondary fetal outcomes included Apgar score and umbilical cord pH.

Associations between maternal BMI categories and outcomes were examined using multivariable regression adjusting for pre-delivery confounders. Addressing decision-making about hospital choice, intrapartum variables are considered intermediate variables and not adjusted for.

Results

Neonates born to women with obesity class II (BMI 35-39.9 kg/m^2) had a 7,64% (95% CI: [4,01% 13,0%]) risk of transfer to a neonatal intensive care unit, compared with 3,84% (95% CI: [3,25% 4,51%]) among neonates of normal-weight women. This difference remained statistically significant in multivariable regression (OR = 1.97 p = 0.035). In contrast, Apgar scores and umbilical cord pH did not differ substantially.

Discussion

In this care level IV cohort, obesity class II was associated with a modestly higher probability of NICU transfer, while immediate neonatal condition was comparable to that of normal-weight pregnancies. To support informed decision-making, pregnant women with BMI 35-40 kg/m^2 should be counselled using absolute risks in addition to relative differences or odds ratios. Given the higher likelihood of neonatal transfer, they may consider delivery in a facility with an on-site neonatal department (care levels I-III) to reduce the chance of postnatal out-of-hospital transfer.

[30-P4-02] Diabetes-related distress and its association with oral health-related quality of life in young individuals with type 1 diabetes

Presenter: Abdur Rasheed

Authors: Abdur Rasheed [1]; Sara Imtiaz [2]

Affiliations: [1] Dow University of Health Sciences, Karachi; [2] Indus Hospital, Karachi

Introduction

Diabetes distress is a significant psychosocial burden among adolescents and young adults with type 1 diabetes, a transitional life stage marked by increasing self-management responsibilities and socio-behavioural challenges. Literature highlights links between psychological, metabolic, and oral health outcomes in T1D. However, data on Diabetes distress prevalence and its relationship with oral health-related quality of life (OHRQoL) are limited in Pakistan. The study aims to identify the proportion of diabetes distress among young individuals with T1D and examine its association with OHRQoL and other clinical and demographic factors.

Methods

A cross-sectional study was conducted among 351 individuals aged 15-24 years attending outpatient clinics at Dow University of Health Sciences (DUHS) and Baqai Institute of Diabetology and Endocrinology (BIDE). Ethical approval for this study was obtained from IRB - BIDE and DUHS. Participants were recruited via convenience sampling and provided written informed consent. Diabetes Distress was assessed using the Diabetes Distress Scale (T1DDS), whereas OHRQoL was assessed using the Oral Health Impact Profile-14. Data on socio-demographics, clinical variables, oral health were also obtained. Univariable and Multivariable logistic regressions identified factors associated with Diabetes Distress, and model performance was evaluated using ROC and calibration plots.

Results

The mean age of the participants was 20.12 $\pm$ 3.10 years with nearly equal gender distribution. Overall, 44% participants reported Diabetes Distress, with feelings of powerlessness and emotional strain most prominent. Poor OHRQoL (AOR 4.53, $p<0.001$) and a history of DKA in the past year (AOR 1.86, $p=0.048$) increased odds of Diabetes Distress, while longer T1D duration (AOR 0.92, $p=0.002$) and dental fluorosis (AOR 0.199, $p=0.046$) were protective. Associations with dental caries, periodontal pockets, and overall oral health were significant in univariate analysis but attenuated after adjustment. ROC analysis indicated good discriminative ability calibration plots showed adequate model fit.

Discussion

Diabetes Distress is highly prevalent among young people with T1D and strongly associated to poor OHRQoL and certain oral health conditions. Integrated psychosocial and oral health support should be prioritized in T1D management during adolescence and early adulthood. Addressing diabetes distress alongside oral health needs is essential to improve overall well-being during the critical transition from adolescence to early adulthood.

[30-P4-03] Evaluating Blood Glucose and Clinical Outcomes in Preterm Infants Using Genetic Instruments

Presenter: Lisa-Marie Nuxoll

Authors: Lisa-Marie Nuxoll [1]; Franziska Hanßmann [2]; Wolfgang Göpel [2]; Inke R. König [1]

Affiliations: [1] Institute of Medical Biometry and Statistics, University of Lübeck, University Hospital Schleswig-Holstein, Campus Lübeck, Lübeck, Germany, Lübeck; [2] Department of Paediatrics and Adolescent Medicine, University of Lübeck, University Hospital Schleswig-Holstein, Campus Lübeck, Lübeck, Germany, Lübeck

Introduction

Maintaining stable blood glucose levels in preterm infants is a clinical challenge, with potential implications for the further development of the child. Furthermore, dysglycaemia is common in preterm infants and may be associated with adverse outcomes, including mortality and severe complications. However, the causal relationships remain unclear due to confounding and reverse causation. We aimed to investigate whether genetically proxied blood glucose levels are associated with mortality or severe complications in preterm infants by leveraging polygenic scores (PGS) derived from adult populations.

Methods

We analyzed longitudinal blood glucose measurements from preterm infants during the first 70 days of life. PGS for glucose-related traits including blood glucose, HbA1c, type 1 and type 2 diabetes, and insulin resistance were calculated using results from previously published genome-wide association studies conducted in adult populations. We assessed associations between these PGS and

measured neonatal blood glucose levels. Additionally, we performed Mendelian randomization (MR) analyses using these genetic instruments to evaluate potential causal effects of blood glucose on mortality and severe complications, applying multiple MR methods to assess robustness.

Results

None of the evaluated PGS showed meaningful associations with blood glucose levels in preterm infants. This suggests limited transferability of adult-derived genetic predictors to neonatal glycemic regulation. In line with this, MR analyses did not provide evidence for an association between genetically proxied adult blood glucose traits and mortality or severe complications in preterm infants across all methods applied.

Discussion

Our findings highlight important limitations in extrapolating genetic instruments derived from adult populations to neonatal settings. The lack of association between PGS and neonatal glucose levels suggests fundamental differences in the underlying biology of glycemic control between preterm infants and adults. Consequently, MR analyses based on these instruments may not be informative. Importantly, these negative results caution against the uncritical application of adult-derived genetic tools in early-life research and underscore the need for neonatal-specific genomic studies. Future work should focus on identifying age-appropriate genetic determinants of glycemic regulation to better understand its role in neonatal outcomes.

[30-P4-04] Population-based study of liver cirrhosis epidemiology in Wales, 2000-2023

Presenter: Kristine Bjarnum Rasmussen

Authors: Kristine Bjarnum Rasmussen [1]; Colin Crooks [1]; Aidan O'Keeffe [1]; Joe West [1]; Peter Jepsen [2]

Affiliations: [1] University of Nottingham, Nottingham; [2] Aarhus University, Aarhus

Introduction

UK liver-related mortality has increased considerably since 1970, especially in cases of alcohol, viral hepatitis, and metabolic syndrome, but the contemporary trends in the occurrence of cirrhosis and survival following diagnosis are unknown. This study evaluated the incidence, prevalence, and one-year survival rates of cirrhosis among Welsh patients.

Methods

This population-based study uses anonymised linked GP, hospital, and death records from SAIL Databank. Welsh GP-registered liver cirrhosis patients diagnosed between 2000 and 2023 were followed until they left Wales, died, or the study ended.

Results

During a follow-up period including 45,118,441 person-years, 21,393 individuals were diagnosed with liver cirrhosis 61.5% were male, and 81% of the cases were individuals over the age of 50. In 2023, there were 331.93 cases per 100,000 individuals. The compensated prevalence ratio was 2.14 (95% CI 2.04-2.24) compared with decompensated cases. The incidence of cirrhosis rose 2.5-fold from 29 (95% CI 26-32) per 100,000 in 2000 to 73 in 2023, irrespective of demographic changes. Compensated cirrhosis (absent ascites or varices) grew by 3.3 times, whereas decompensated cirrhosis rose by only 1.7 times. The incidence of alcohol-related cirrhosis increased twofold, but metabolically dysfunction-associated steatotic liver disease escalated 3.6 times. The yearly mortality risk for compensated patients decreased slightly (OR = 0.97), whereas that for decompensated patients remained unchanged (OR = 1.00). The incidence temporal profile for MASLD increased notably over the period, and IMD2 to IM5 increased significantly relative to IMD1 however, none of these factors is statistically significant when evaluating 1-year survival probabilities.

Discussion

The Welsh cirrhosis population between 2000 and 2023 has increased, with older patients, more cases of early disease severity, and MASLD cases. The one-year survival probabilities for compensated cirrhosis improved, while those for decompensated cirrhosis were unchanged. The 2015 National Welsh guidelines, which encourage active case finding in obese or type 2 diabetic adults, may explain these findings.

[30-P4-05] Using a large-scale post-market survey to derive information about audio processor use and satisfaction in different CI user groups

Presenter: Dorothée B. Hoppe

Authors: Dorothée B. Hoppe [1]; Mareike Billinger-Finke [1]; Elena Grandelis [1]; Karin Koinig [1]; Patrick F. Connolly [1]; Ilona Anderson [1]

Affiliations: [1] MED-EL Elektromedizinische Geräte GmbH, Innsbruck

Introduction

Cochlear implants (CIs) are used in combination with external audio processors which convert sound into signals that enable CI users to perceive auditory information. Tracking users' satisfaction with audio processors is crucial for manufacturers to monitor product feedback, but can also offer insights to clinicians about how different user groups respond to different device characteristics.

The Audio Processor Satisfaction Questionnaire (APSQ) is a validated questionnaire running for over 5 years as a post-market survey. It provides observational data on processor satisfaction and daily usage, and on demographic and device characteristics. Aim of the present analysis was to establish a reference cohort for satisfaction scores and to investigate device use and satisfaction across user groups to identify groups at higher risk of experiencing device-related problems.

Methods

In this retrospective analysis, responses of 861 CI users were analyzed. Satisfaction scores (continuous 0-10 total score and three subscales) were modeled with multivariable quantile regression daily usage (0-3h, 3-6h, 6-9h, 9-12h, >12h) was modeled with multivariable ordinal logistic regression. All models included the predictors age, gender, audio processor type (behind-the-ear, off-the-ear), and a combination of indication and hearing solution. Satisfaction score models were used to derive population norms.

Results

Satisfaction scores were generally high for all user groups and device types (median (IQR) = 8.7 (7.9, 9.3) points). A median regression model suggested only small differences in satisfaction scores between different user and device characteristics (all marginal effects 12 h/day) and ordinal logistic regression models suggested that a) higher device satisfaction (1-point increase) was associated with longer daily usage (inverse OR = 1.44 95% CI: 1.27-1.63 p<.001) and b) CI users with single-sided-deafness show shorter daily usage than bilateral CI users (OR = 2.78 95% CI: 1.81-4.27 p<.001).

Discussion

Our results suggest that differences in audio processor satisfaction between user groups and processor types are small. However, we find that reported daily usage increased with increased satisfaction and decreased for specific user groups. Identifying factors associated with reduced device usage may enable better targeted support in hearing rehabilitation after cochlear implantation.

[30-P4-06] Identifying hidden heterogeneity in district-level anaemia determinants among women of reproductive age in India: A Latent Class Analysis of NFHS-5 data

Presenter: Roshni Tony

Authors: Roshni Tony [1]; Thennarasu Kandavel [2]; Nandakumar Dalavaikodihalli Nanjaiah [2]

Affiliations: [1] National Institute of Mental Health and Neurosciences, Bangalore; [2] National Institute of Mental Health and Neurosciences, Bengaluru

Introduction

Anaemia is a condition affecting red blood cells or haemoglobin concentration, mainly appearing in women of reproductive age (WRA) (15-49 years). Anaemia is predominant in low and middle-income countries such as India, with district-level variations driven by differences in healthcare access, socio-economic development and lifestyle behaviours. Since multiple determinants act simultaneously, examining indicators individually conceals underlying patterns, emphasising the need to identify latent district subgroups based on multiple anaemia-related indicators. The objectives are to identify latent district classes that reflect varying levels of vulnerability associated with anaemia among WRA, and validate the identified classes by examining whether external characteristics such as women's education, contraceptive use, selected health and behavioural risk factors differ across district classes.

Methods

Data from NFHS-5 (2019-2021) district factsheets were used, covering 703 districts. Seven anaemia-related health indicators were selected. Quartile-based categorisation of continuous variables was obtained to perform Latent Class Analysis (LCA). Models with varying numbers of classes (2 to 6) were compared using fit indices. Six external variables (in percentages), related to women's education, contraceptive use, health and behavioural risk factors, were used to validate classes using Kruskal-Wallis and Bonferroni-adjusted post-hoc tests.

Results

Four latent district classes were identified: Class 1-High maternal healthcare access (n=262, 37%), Class 2-Low maternal healthcare access (n=109, 16%), Class 3-Low risk (n=182, 26%) and Class 4-Vulnerable districts (n=150, 21%). External variables differed significantly across classes, indicating meaningful variation in anaemia-related indicators. Female literacy was highest in Class 3 and lowest in Class 4. Contraceptive pill use was highest in Class 2, while high blood sugar and high blood pressure were highest in Class 3. Tobacco use was highest in Class 4, and alcohol consumption was low overall but varied across classes.

Discussion

The findings reveal heterogeneity in district-level conditions related to anaemia among WRA. Differences in literacy, reproductive health behaviour, lifestyle, and healthcare access indicate that multiple determinants act together in shaping district health profiles. Classifying districts into latent subgroups supports targeted interventions, enabling planning and implementation of health strategies at district level. In conclusion, joint assessment of determinants identifies four distinct classes that help decide health strategies at district level.

[30-P4-07] Age-specific patterns in prevalent colorectal cancer (CRC) and associated dietary factors: Results from the German National Cohort (NAKO)

Presenter: Virginie Stanislas

Authors: Virginie Stanislas [1]; Stefanie Jaskulski [1]; Peggy Sekula [1]; Verena Katzke [2]; Gianni Panagiotou [3, 4]; Karin B. Michels [1]

Affiliations: [1] Institute for Epidemiology and Prevention, Medical Center - University of Freiburg, Freiburg., Freiburg im Breisgau; [2] Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg; [3] Institute of Microbiology, Friedrich Schiller University, Jena., Jena; [4] Department of Microbiome Dynamics, Leibniz-HKI, Jena., Jena

Introduction

The incidence of early-onset colorectal cancer (EO-CRC, <50 years) has increased markedly in recent decades and tends to be more aggressive than late-onset CRC (LO-CRC). Its etiology remains unclear and is likely multifactorial. Generational lifestyle changes, including rising obesity, dietary

shifts and reduced physical activity, may contribute. This study investigates age-specific CRC patterns across anthropometrics, lifestyle, clinical, and dietary factors.

Methods

We analyzed baseline data (2014-2019) from the NAKO and German cancer registries to identify prevalent first primary CRC cases (ICD-10 C18-C20) diagnosed up to 10 years before baseline and incident cases diagnosed during follow up (up to 9 years). Among 204,676 participants 322 prevalent (70 EO-CRC, 252 LO-CRC) and 353 incident CRC cases (37 EO-CRC, 316 LO-CRC) were identified.

Analyses focused on established CRC risk factors and nutritional patterns derived using principal component analysis (PCA). For prevalent CRC, controls were matched 4:1 on age and sex, and conditional logistic regression models were used to estimate odds ratios (OR). For incident CRC, Cox models with age as the time scale were used to estimate hazard ratios (HR). Separate models were used for EO-CRC and LO-CRC.

Results

Preliminary analyses showed that incident LO-CRC risk was lower in females (HR 0.71 [0.55, 0.91]) and higher for former smokers (HR 1.34 [1.02, 1.75]). Although not significant, trends suggested increased risks with metabolic factors and a high overall intake pattern (PC1), and decreased risks with sedentary behavior and a plant-based versus meat/processed-food pattern (PC2). No clear associations were observed for EO-CRC, possibly due to limited case numbers. However, trends suggested increased risks with metabolic factors and decreased risks for dietary patterns (PC1 and PC2). Prevalent EO-CRC was significantly associated with higher body fat and elevated triglycerides, and lower smoking compared to EO-controls. In contrast, prevalent LO-CRC was associated with lower alcohol intake, smoking, sedentary time, and hyperlipidemia likely reflecting post-diagnosis changes.

Discussion

Our results indicate potential heterogeneity in risk profiles between EO-CRC and LO-CRC. While established factors were associated with LO-CRC, evidence for EO-CRC remained inconclusive. Ongoing analyses aim to investigate dietary exposures in greater detail and to assess additional risk factors.

[30-P4-08] Modeling bimodality of diagnosis age in cutaneous melanoma

Presenter: Anna Oksanen

Authors: Anna Oksanen [1]; Matteo Bottai [2]; Kari Nielsen [3, 4, 5]; Hanna Eriksson [6, 7]; Elisavet Syriopoulou [1]; Therese Andersson [1]

Affiliations: [1] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm; [2] Institute of Environmental Medicine, Karolinska Institutet, Stockholm; [3] Department of Dermatology, Sk˚ane University Hospital, Lund; [4] Department of Clinical Sciences, Dermatology, Lund University Skin Cancer Research Group, and Lund Melanoma Study Group, Lund University, Lund; [5] Department of Dermatology, Helsingborg Hospital, Helsingborg; [6] Department of Oncology and Pathology, Karolinska Institutet, Stockholm; [7] Theme Cancer, Unit of Head-Neck-, Lung- and Skin Cancer, Skin Cancer Center, Karolinska University Hospital, Stockholm

Introduction

The age-specific incidence of cutaneous melanoma has traditionally not been considered to present bimodality, yet indications of it exist especially for females. However, this bimodality has not been much studied. We used a novel method to investigate potential bimodality in the age distribution of cutaneous melanoma. As age is often recorded in whole years, most standard estimators of age distributions have an in-built downward bias. Our proposed method takes into consideration the interval censored nature of age measurement in a flexible way, without distributional assumptions.

Methods

Data was acquired from the Swedish Melanoma Register and included everyone in Sweden diagnosed with cutaneous melanoma from 2000 to 2014 (35 699 individuals). We fitted a unimodal and bimodal model to the age distribution, comparisons were made using likelihood ratio test and AIC. The bimodal model utilized a latent structure as the existence of two separate groups was not known. The cumulative distribution at lower and upper age-bounds was based on the probability integral transform and approximated using Taylor expansion. Estimation was done using maximum likelihood. The likelihood function was formulated as the difference of the upper and lower bounds. Optimisation was done in R using the nlm -function.

Results

When compared using the likelihood ratio test, the age distribution of melanoma incidence is better explained by using a mixture model (p-value <= .000), than a univariate model of the same degree of flexibility. However, when comparison is made using AIC, the adjustment for the additional parameters makes the univariate model appear superior (AIC -17.9 and -11.8 for the univariate and bivariate respectively). A higher order expansion allows for bimodality, significantly (p-value <= .000) improving the univariate model, but not enough to push the AIC under that of the simplest model.

Discussion

In large data sets the estimation of whole distributions efficiently and without bias is a relevant, but often overlooked, question. Our example introduces a new way of approaching the problem, utilizing the interval censored nature of data measurement. We investigated the distribution of age at diagnosis for cutaneous melanoma observing the degree of Taylor polynomials required to approximate the distribution efficiently.

[30-P4-09] Sex-specific incidence of thyroid cancer in districts with and without a nuclear power plant in Germany from 2000 to 2023

Presenter: Emilio Gianicolo

Authors: Henriette Dittrich [1]; Maria Blettner [2]; Klaus Kraywinkel [3]; Emilio Gianicolo [1]

Affiliations: [1] Institut für Medizinische Biometrie, Epidemiologie und Informatik der Universitätsmedizin der Johannes Gutemberg-Universität, Mainz; [2] Universitätsmedizin Mainz, Mainz; [3] Robert Koch-Institut, Abteilung für Epidemiologie und Gesundheitsmonitoring, Berlin

Introduction

Incidence rates for thyroid cancer have increased worldwide, partially attributable to widespread screening and novel diagnostic tools (1). Ionizing radiation, classified by IARC in group 1 (2), is an established risk factor for thyroid cancer and elevated incidence rates have been reported following the reactor accident at Chernobyl (3, 4). However, several studies investigated thyroid cancer mortality and incidence among people living close to nuclear power plants (NPP) with conflicting results (5-10). This study compares incidence rates of thyroid cancer in German districts with and without an NPP.

Methods

We used an ecological study design to investigate the incidence of thyroid cancer in persons aged ≥ 20 years living in all 400 German districts (Landkreise) from 2000 to 2023, classifying those with at least one NPP within their boundaries as 'districts with NPP". Data, stratified by sex, 10-year age-groups, year of diagnosis, and district, were provided by the Zentrum-für-Krebsdaten. We calculated age standardized incidence ratios (SIR) using districts without an NPP as the reference. Furthermore, we fitted a mixed Poisson regression model, with a random effect at federal state level using log-population as offset.

Results

We observed 36 140 male and 85 982 female cases. In men, we found crude rates (per 100 000) of 4.9 in districts with an NPP and 5.4 in districts without. In women, the rates were 10.9 in districts with an NPP and 12.2 in districts without. The SIR for men was 0.89 (95% CI: 0.84-0.93) and 0.88 (95% CI: 0.85-0.91) in women, indicating lower incidence in districts with an NPP. Poisson regression results were consistent with the SIR-analysis.

Discussion

This study found no increased incidence of thyroid cancer in adults in districts with an NPP compared to districts without. Key limitations include lack of dose estimation and confounder adjustment, and ecological fallacy. Strengths include high quality cancer data and consistent results across methodological approaches. Future studies might use data at smaller administrative levels as well as inverse-distance weighted metrics as proxy of potential exposure to NPPs (9). Additionally, adjusting for potential confounders such as socioeconomic position could further improve the validity.

[30-P4-10] Mortality among former employees in German uranium processing and underground mining

Presenter: Veronika Deffner

Authors: Veronika Deffner [1]; Michaela Kreuzer [1]; Beate Hochstrat [1]; Peter Scholz-Kreisel [1]; Nora Fenske [1]

Affiliations: [1] Federal Office for Radiation Protection, München (Neuherberg)

Introduction

Former uranium miners and processing workers at the German Wismut company were exposed to various types and levels of ionising radiation and dust. We aim to investigate the mortality rates of Wismut processing workers in relation to the general male population, and compare our findings with those of underground miners.

Methods

We studied a sub-cohort of 4,161 male processing workers, who were employed in uranium processing at the Wismut for at least 180 days between 1946 and 1989, and who had never worked in underground or open-pit mining. This sub-cohort was analysed over a period from 1960 to 2018, and compared to underground miners. Observed deaths (O) were evaluated against expected deaths (E), based on external mortality rates for the general male population, using standardized mortality ratios (SMR = O/E). Exposure to radon progeny, external gamma radiation, and long-lived radionuclides was retrospectively estimated using job-exposure matrices, from which total absorbed organ doses were calculated.

Results

The highest cumulative organ doses were received by the lungs (median=53 milligray (mGy) 25% quantile=16 mGy 75% quantile=134 mGy), mainly due to radon progeny, and by the bone surface (median=20 mGy 25% quantile=6 mGy 75% quantile=51 mGy), primarily due to gamma radiation. Processing workers had lower all-cause mortality (O=2,385 SMR=0.86 95% confidence interval (CI): 0.83-0.90), all-cancer mortality (O=671 SMR=0.94 0.87-1.01), and mortality from most specific causes, compared to the general population. Unlike underground miners, there were no significant increases in SMRs for lung cancer (O=212 SMR=1.04 95% CI: 0.90-1.19) and non-malignant respiratory diseases (O=161 SMR=0.82 95% CI: 0.70-0.96). Among cancers, only SMRs of central nervous system cancers (O=25 SMR=1.54 95% CI: 1.00-2.27) and non-Hodgkin lymphoma (O=18 SMR=1.15 0.68-1.82) were elevated. Consistent with underground miners, there was a significant excess mortality from cerebrovascular diseases (O=229 SMR=1.26 95% CI: 1.10-1.43).

Discussion

While increased SMRs were observed among underground miners for lung cancer, silicosis and some other diseases, no clear excess for any cause of death, except for cerebrovascular diseases, was present among uranium processing workers. Further research on processing workers will focus on

dose-response analyses within the International Pooled Analysis of Uranium Processing Workers (iPAUW).

[30-P4-11] Socioeconomic deprivation and health-related quality of life in kidney and urinary tract cancer: an ecological analysis of national survey data

Presenter: Felix Keller

Authors: Franziska Gross [1]; Johannes Giesinger [1]; Felix Keller [2]

Affiliations: [1] University Hospital of Psychiatry II, Medical University of Innsbruck, Innsbruck; [2] Department of Internal Medicine IV, Nephrology & Hypertension, Medical University of Innsbruck, Innsbruck

Introduction

Health-related quality of life (HRQoL) is an important outcome for kidney and urinary tract cancer survivors, where symptoms and functional limitations often persist. Socioeconomic deprivation is associated with adverse cancer outcomes, but its relationship with HRQoL at the population level remains incompletely characterised. Routinely published area-level data from a national survey enable comparisons of patient-reported HRQoL 18 months post cancer diagnosis across regions in England.

Methods

We analysed publicly released Integrated Care Board (ICB)-level data from the NHS Cancer Quality of Life Survey (England, June 2025) for kidney and urinary tract cancer (42 ICBs 7,695 patients contributing to the published aggregates). Overall and domain-specific HRQoL were assessed using the EORTC QLQ-C30 questionnaire.

Socioeconomic deprivation was captured using area-level deprivation quintiles (Index of Multiple Deprivation 2019) and summarised into a deprivation imbalance index, defined as a log-ratio contrasting the relative concentration of patients in more versus less deprived groups.

We fitted inverse-variance weighted multivariable mixed models to account for differences in precision between ICBs, adjusting for ICB-level age, gender, and disease stage, with fixed effects for NHS region and random intercepts for Cancer Alliance.

Results

Greater deprivation imbalance was associated with lower overall HRQoL (β = -1.79 per index unit 95% CI -2.63 to -0.96), independent of demographic and clinical factors. Comparing regions with low versus high deprivation (10th vs 90th percentile of the index) corresponded to an approximately 3.5 points lower QLQ-C30 Summary Score (95% CI -5.1 to -1.9).

Across domains, associations were consistently negative, with the largest effects observed for insomnia, fatigue, and role functioning (approximately -4 to -6 points).

Discussion

These findings indicate lower average HRQoL in regions where patient populations are more socioeconomically disadvantaged, even after accounting for case-mix and geographic structure. The consistency across multiple domains suggests a general rather than domain-specific pattern of disadvantage, with implications for monitoring inequalities and targeting supportive care in kidney and urinary tract cancer survivors at the system level.

As an ecological analysis, results reflect differences between populations and areas rather than individual-level effects, but the robustness of findings supports further investigation using linked individual-level data.

[30-P4-12] Examples of bias in data from translational research

Presenter: Ella Maria Jentsch

Authors: Ella Maria Jentsch [1]; Ulrich Sax [1]; Jonas Hügel [2, 1, 3, 4]; Sophia Rheinländer [1]; Nils Beyer [1]; Elisabeth Hessmann [5]

Affiliations: [1] University Medical Center Göttingen, Department of Medical Informatics, Göttingen; [2] Department of Health Technologies, Technical University of Denmark, Kongens Lyngby; [3] University of Göttingen, Campus Institut Data Science, Section of Medical Data Science, Göttingen; [4] Danish Center of Sleep Medicine, Rigshospitalet, Glostrup; [5] University Medical Center Göttingen, Clinic for Gastroenterology, Gastrointestinal Oncology and Endocrinology, Göttingen

Introduction

Data from electronic health records are increasingly important for research and the development of data-driven decision support systems in healthcare digitization. However, these data sources carry the risk of systematic bias reflecting existing social inequalities, which may be reinforced through artificial intelligence (AI) [1]. This work aims to systematize key bias categories in health data, examine them using a translational research dataset, and analyze the role of domain and process knowledge in identifying biases.

Methods

A structured literature review was conducted to identify overarching bias categories in health data. Based on this review, three main categories were defined: data bias, algorithmic bias, and societal bias. These were conceptualized as interconnected elements forming a mutually reinforcing cycle. Additionally, a retrospective analysis of the MolPAC dataset, derived from a clinical registry of pancreatic cancer patients, was performed [2]. Potential representation biases were examined with respect to insurance status, population density, distance to treatment location, age, and gender.

Results

The analysis revealed a statistically significant difference in appointment frequency between patients with statutory and private insurance. Furthermore, a significant negative correlation was found

between the distance of residence and the number of documented appointments. In contrast, population density, gender, and age showed little to no significant effects. These findings suggest that structural access conditions, particularly insurance type and geographic distance, influence the visibility of patients in clinical datasets.

Discussion

The results demonstrate that bias identification and evaluation should not rely solely on statistical methods. Incorporating domain and process knowledge is essential for interpreting data within its real-world context. Bias must be understood as a socio-technical phenomenon that spans technical, institutional, and societal levels. Overall, this work highlights that translational health data can contain inherent biases that must be considered in the development of fair and robust AI systems. Systematic bias reflection is therefore a crucial prerequisite for the equitable and evidence-based use of clinical data.

[1] D. Leslie et al., "Does "AI" stand for augmenting inequality in the era of covid-19 healthcare?", BMJ, March 2021, doi: 10.1136/bmj.n304.

[2] "MolPac". [Online]. https://gccc.umg.eu/kfo5002/blog-kfo/molpac/

[30-P4-13] Data Deserts in Colorectal Cancer with Structured Missingness and Representation Gaps as Pre-modelling Fairness Problems in the UK Biobank

Presenter: Corinna Meeßen

Authors: Corinna Meeßen [1]; Carolin Schneider [1]

Affiliations: [1] Uniklinik RWTH Aachen, Aachen

Introduction

Population-scale cohorts support thousands of cancer studies, yet conclusions are often assumed generalizable despite systematic differences in data completeness, measurement, and subgroup representation. Fairness evaluations focus on model outputs, despite upstream data quality differences that constrain achievable performance across subgroups. We map data deserts and quantify subgroup bias prior to modeling.

Methods

10,264 colorectal cancer (C18-C21) participants from UK Biobank were analyzed across five predictor panels: demographics, lifestyle, anthropometrics, vitals, and laboratory biomarkers. Missingness was quantified as proportions using Wilson 95% confidence intervals. Subgroup categories (7xethnic groups, 5xdeprivation quintiles, 2xsex, 5xagegroups) yielded 95 subgroup×panel combinations. Data deserts were defined as elevated missingness above the 'White' reference rate, $n<100$, or both. Stratification across all dimensions was assessed using ANOVA. Logistic regression evaluated predictors of missing biomarkers with prespecified references. Missingness mechanisms were inferred using convergent evidence.

Results

Overall, 23.1% of participants had ≥1 missing biomarker. Missingness ranged from 18.2%-38.0% across ethnic subgroups (Chinese n=22 excluded from adjusted analysis). No significant correlation with subgroup size was observed (r=0.04, p=0.94, limited power given k=7). Of the 95 subgroup-panel combinations, 67% met ≥1 data desert criterion, 15% met both, all within the the non-White ethnic subgroups. Enrichment of missingness by non-White ethnicity was observed across the laboratory, anthropometrics and vitals panels ($p<0.05$) except for the Chinese group (n=22). Higher deprivation was associated with greater missingness in lifestyle and anthropometric panels ($p<0.05$). After adjustment, the 'Other' ethnic group had higher odds of missing biomarkers (OR=1.97, 95%CI=1.11-3.50, p=0.021), while males had lower odds (OR=0.80, 95%CI=0.73-0.87, $p<0.001$). All five non-White subgroups had n<100, indicating representation gaps limiting statistical power but not explaining the observed missingness patterns. Evidence was consistent with missing-at-random conditional on observed demographics.

Discussion

Data completeness is systematically structured by subgroup, creating data deserts that co-occur with representation gaps and are not explained by sample size alone. These patterns suggest achievable model performance may be constrained across subgroups before model development. Ongoing analyses link these patterns to calibration error and clinical decision impact at prespecified time horizons, supporting quality-aware and equitable cancer risk modeling.

[30-P4-14] Cognitive state and fine motor function in the CHRIS study: scope for missingness and selection

Presenter: Mousumi Banerjee

Authors: Roberto Melotti [1]; Aaron Kaat [2]; Martin Gögele [1]; Mousumi Banerjee [3, 1]; Peter Pramstaller [1]; Richard Gershon [2]; Cristian Pattaro [1]; Stella Wang [2, 1]

Affiliations: [1] Eurac Research, Bolzano; [2] Northwestern University, Chicago; [3] University of Michigan, Ann Arbor

Introduction

Fine motor performance involves both cognitive and motor control neurocircuitry and may represent an early marker of cognitive impairment and dementia risk. These measurements in community studies are hampered by self-selection of participants and missing data, possibly affecting prevalence and association estimates. We used data from the Cooperative Health Research in South Tyrol (CHRIS) study to assess global cognitive function and its relationship with spiral drawing performance, accounting for selection and missingness.

Methods

Out of 13,367 CHRIS participants with available data, we included 8,230 (61.6%) without tremor related conditions who completed 6 replicates of the digital spiral drawing test. Among these, 7,401

had full 30-item response (CC), 216 had only partial item response, and 613 had full missing items to the Mini-Mental State Examination (MMSE). Spiral drawing performance was measured by multiple spatial-time metrics, including mean drawing error from the smoothed trace and median speed, averaged across 6 replicates. Post-stratification weighted multiple imputation (MI) of single MMSE missing items and other imputed covariates was applied.

Prevalence of MMSE classification and adjusted generalised linear models for the association between MMSE score and drawing performance were estimated in CC and MI data for comparison.

Results

The proportions of 7,401 participants with complete data grouped by MMSE scores [30,29-27,26-24, <24] were 53.4%, 36.2%, 9.4% and 1.0%, respectively. Corresponding weighted figures for the lowest two MMSE groups with complete data increased to 12.0% and 1.7% and to 12.3% and 2.2% among 8,102 participants with imputed data, respectively.

MMSE score had an inverse linear association with either spiral metric in all analyses. There was evidence of an association between mean drawing error and inverse MMSE score in both CC [coeff.=0.008 (95%CI=0.003, 0.014)] and MI [0.008(0.004, 0.011)] data. However, evidence of an association between median speed and MMSE was moderate in CC [-0.623(-1.115, -0.132)] and null in MI [-0.188(-0.645, 0.269)].

Discussion

Prevalence estimates of low cognitive function were affected by selection and missingness in the CHRIS study. However, associations with fine motor performance were consistent for trace error and inconsistent for speed between CC and MI analyses, suggesting different patterns of association bias by metrics.

[30-P4-15] Measurement error in Chronic Pain

Presenter: Aleksandra Turkiewicz

Authors: Aleksandra Turkiewicz [1]; Marius Henriksen [2]; Elisabeth Bandak [2]; Elisabeth Ginnerup-Nielsen [2]; Martin Englund [1]

Affiliations: [1] Clinical Epidemiology Unit, Orthopedics, Clinical Sciences Lund, Lund University, Lund, Sweden, Lund; [2] The Parker Institute, Bispebjerg and Frederiksberg Hospital, Copenhagen, Denmark, Copenhagen

Introduction

Visual Analogue Scale (VAS, 0-100 mm) remains the most used tool for assessment of pain in musculoskeletal conditions being the leading cause of disability. This and similar self-reported scales suffer from large measurement error that leads to regression to the mean in trials, and confusion about the size of clinically relevant changes. The aim of this work is to estimate the extent to which measurement error can be minimized by using the average of repeated measurements.

Methods

We used data from 101 participants from both groups of a randomized clinical trial including patients with knee osteoarthritis comparing two types of conversations in addition to saline injection (i.e. placebo).1 Daily measurements of VAS pain from day -6 (6 days pre-randomization) to day 13 were used. The pre- and post-randomization data were used separately. The participants' state can be assumed to be stable in the 6 days pre-randomization and 12 days post randomization. We defined two pre-randomization periods, each with a duration of 3 days and with up to 3 measurements per period. For post-randomization data, we defined two periods of duration 6 days with up to 6 measurements per period. We estimated standard error of measurement (SEM) and smallest detectable change (SDC) for the two periods using average from 1 to 3 (pre-randomization) or 1 to 6 (post-randomization) measurements per period.

Results

The SEM using two 3-day periods was 12.0 (95%CI 10.4, 13.7) when using one measurement and 8.9 (7.6, 10.1) or 7.5 (6.5, 8.5) when using average of two or three measurements. The SEM using two 6-day periods was 12.2 (10.5, 13.9) when using one measurement and 9.7, 9.1, 9.1, 9.0 and 8.6 when using averages of two to six measurements. The corresponding SDCs where 33.9 for one measurement and 23.8 when using average of 6 measurements.

Discussion

A simple and feasible pain assessment using average of at least three repeated measurements can lead to decrease of SEM and SDC by 25% to 35%. However, the SDC is large (~24 points on a 0 to 100 scale) even when using average of six measurements.

1. DOI: 10.1016/j.joca.2024.07.005

[30-P4-16] The Wilcoxon-Mann-Whitney test with incomplete data: mutiple imputation for the Hodges-Lehmann estimator

Presenter: David Petroff

Authors: David Petroff [1]; Miroslav Bacak [1]

Affiliations: [1] Clinical Trial Centre, Leipzig University, Leipzig

Introduction

Missing data is a common problem in RCTs and one popular approach to overcome the related difficulties is the multiple imputation method, as introduced by Rubin [1]. Whereas this method is well-established for parametric models, it has not been studied in the context of non-parametric statistics. We close this gap by deriving an imputation method for the Hodges-Lehmann estimator [2] which is a standard tool in non-parametric statistics for estimating the location parameter associated with the Wilcoxon-Mann-Whitney test.

Methods

We follow the standard approach for deriving Rubin's rules in the parametric case and extend it to the non-parametric one. Having obtained an imputation method, we study its properties in simulation studies, both in terms of the effects of increasing the number of multiple imputations and comparing the Hodges-Lehmann estimator with an estimator from linear regression models, where multiple imputations are standard.

Results

Based on a consistent estimator for the asymptotic standard deviation of the point estimator, we verify that Rubin's rules can be applied in the usual fashion and obtain the desired multiple imputation method for the Hodges-Lehmann estimator. Preliminary results from simulations suggest that multiple imputation for the Hodges-Lehmann estimator is similar to the well-known case from linear regression.

Discussion

A gap has been closed in non-parametric methodology and the Wilcoxon-Mann-Whitney test with its Hodges-Lehmann estimator is now amenable to multiple imputation.

[1] D. B. Rubin, Multiple imputation for nonresponse in surveys, Wiley Series in Probability and Mathematical Statistics: Applied

Probability and Statistics, John Wiley & Sons, Inc., New York, 1987.

[2] J. L. Hodges, Jr. and E. L. Lehmann, Estimates of location based on rank tests, Ann. Math. Statist., 34 (1963), pp. 598-611.

[30-P4-17] The robustness of Rubin's rules and an approximate Bayesian pooling method under posterior skewness

Presenter: Daniel J Phillips

Authors: Daniel J Phillips [1]; Maria D Christodoulou [1]; David Steinsaltz [1]

Affiliations: [1] University of Oxford, Oxford

Introduction

Multiple imputation (MI) is a popular method to deal with missing data in clinical research. Statistical inference after MI is typically combined using a pooling method known as Rubin's rules. Statistical analysis of some clinical datasets with missing data produces skewed posterior distributions, whereas Rubin's rules assume the observed-data posterior is symmetric, which may lead to biased inference. We consider the robustness of Rubin's pooling rules under posterior skewness.

Methods

We present an approximate Bayesian pooling scheme (ABpool) that allows for posterior skewness, while requiring substantially lower computational cost than full Bayesian inference after MI. Specifically, we sample once from a t or Gaussian approximation of the completed-data posterior distribution for each imputed dataset, and pool the samples. We mathematically derive the error in applying Rubin's rules where key assumptions do not hold. We compare the pooling methods in a simulation study with skewed data. We apply both pooling methods to three datasets: to estimate the correlation between viral infection and cervical cancer to predict the risk of death among seriously ill patients and to fit a joint model for antibody decay and COVID-19 vaccine efficacy.

Results

Results from Rubin's rules and ABpool tend to converge as the sample size increases. ABpool gives higher power for logistic regression with a skewed covariate. For linear and Cox regression, Rubin's rules give higher power to detect an effect when the proportion of missing data is high and sample size is moderate, but lower power in small samples when the proportion of missing data is low. Rubin's rules are unable to account for skewness when applied to the clinical datasets, whereas ABpool recreates the results from Bayesian inference after MI.

Discussion

Rubin's rules assume the posterior distribution is symmetric, which does not always hold. Posterior asymmetry can be detected by examining a normal QQ-plot of the maximum likelihood estimates for the multiply imputed datasets. Rubin's rules appear to give robust inferences in simulated data, however ABpool is more efficient in some scenarios. When applied to medical data, ABpool is able to account for posterior skewness which Rubin's rules cannot.

P5 - Time-to-event and Longitudinal Methods

Poster area: P5

[30-P5-01] Random Survival Forest and Tree Analysis Reveal Structured Heterogeneity in Mortality Risk in People with Type 2 Diabetes and Chronic Kidney Disease

Presenter: Monica Kundu

Authors: Monica Kundu [1]; Ash Routen [1]; Kamlesh Khunti [2]; Safoora Gharibzadeh [2]

Affiliations: [1] Diabetes Research Centre, Leicester General Hospital, University of Leicester, Leicester, UK, Leicester; [2] Leicester Real World Evidence Unit, Diabetes Research Centre, Leicester General Hospital, University of Leicester, Leicester, UK, Leicester

Introduction

Understanding how mortality risk is structured among individuals with type 2 diabetes (T2DM) and chronic kidney disease (CKD) is essential for effective risk stratification, yet heterogeneity within this population remains poorly characterised. Traditional approaches often assume homogeneous risk or rely on linear models that fail to capture complex interactions between predictors. We applied machine learning methods to examine how interactions between clinical factors shape mortality risk.

Methods

An observational study was conducted using data on 15,230 adults (mean age 70.9 years, 53.6% female, 95.6% White ethnicity) with T2DM and CKD from the Clinical Practice Research Datalink (CPRD), England. A random survival forest model (100 trees) was used to identify predictors of all-cause and CVD mortality, with variable importance assessed using permutation-based relative importance scores. Model performance was optimised using internal 10-fold cross-validation. A survival tree was applied to examine hierarchical interactions between predictors and to characterise how combinations of clinical factors shape mortality risk.

Results

Over a median follow-up of 7.3 years, 6,290 deaths occurred. Age was identified as the dominant predictor for all-cause and CVD mortality, followed by eGFR, Multiple Long-Term Conditions (MLTCs), and HbA1c. The survival tree identified a hierarchical interaction structure, with age 76 years as the primary split. Among those aged <66 years, absence of MLTCs defined the lowest-risk subgroup (21.6%), while MLTCs with HbA1c ≥5.9 increased mortality to 40.2%. In individuals aged 66-76 years, mortality rose from 33.5% to 46.9% with combined low eGFR (<54) and MLTCs. Among those aged ≥76 years, mortality was primarily driven by age and kidney function, with minimal additional contribution from MLTCs. For CVD mortality, the first split occurred at 74 years. Mortality ranged from 4.3% in individuals aged <66 years with eGFR ≥49 and lower smoking exposure to 28.1% in those aged ≥90 years.

Discussion

Mortality risk is structured by interactions between age, kidney function, MLTCs, and glycaemic status rather than single predictors. Machine learning-based survival approaches provide a transparent framework to identify clinically actionable risk profiles across large routine healthcare datasets, supporting targeted intervention and efficient resource allocation.

[30-P5-02] Comparison of Baseline and Trajectory Information for Predicting Mortality Risk of Acute Myocardial Infarction

Presenter: Keita Fukai

Authors: Keita Fukai [1]; Ayano Takeuchi [1]

Affiliations: [1] Chuo University, Graduate School of Science and Engineering, Tokyo

Introduction

Cardiovascular diseases are among the leading causes of death in Japan, and acute myocardial infarction (AMI) accounts for a substantial proportion of cardiovascular mortality, representing a significant disease burden. In long-term cohort studies based on health check-up data, individuals are at risk of experiencing multiple types of death, necessitating the use of competing risk analysis. Conventional competing risk methods generally assume linear effects however, cardiovascular risk factors may exhibit nonlinear relationships and threshold-like behaviors. Recently, competing risk analysis using interpretable machine learning approaches, such as Neural Additive Models, has gained attention due to its ability to capture nonlinear effects while maintaining interpretability and predictive performance. However, conventional risk scoring primarily relies on static baseline measurements, failing to account for the dynamic, long-term temporal evolution of risk factors. Therefore, this study aims to extract trajectory patterns of risk factors from longitudinal data and incorporate these temporal features into CRISP-NAM to improve predictive accuracy and interpretability.

Methods

In this study, we used long-term follow-up data from the Ibaraki Prefectural Health Study to perform competing risk analysis with AMI mortality as the outcome. First, a prediction model was constructed using CRISP-NAM with baseline clinical variables as predictors. Longitudinal measurements were utilized to identify latent trajectory patterns for each risk factor. Specifically, individuals were stratified into phenotypic trajectory groups using [e.g., Latent Class Growth Analysis / Group-Based Trajectory Modeling]. These trajectory groups were then incorporated as categorical predictors to construct an additional CRISP-NAM model. The incremental predictive value of trajectory features was evaluated by comparing the CRISP-NAM baseline model with an augmented model incorporating longitudinal features, focusing on discrimination (C-index) and calibration (Brier score).

Results

The results will be presented at the conference. Through this comparison, we demonstrate the utility of integrating temporal trajectories into deep learning-based competing risk models to improve both clinical risk stratification and mechanistic interpretability.

Discussion

The discussion will be presented at the conference.

[30-P5-03] Improving Competing Risks Prediction Using Random Survival Forests with Multiple Imputation and Covariate-Dependent Censoring.

Presenter: Moss. Maharun Nahar Jui

Authors: Moss. Maharun Nahar Jui [1]; Md Hasinur Rahaman Khan [2]

Affiliations: [1] Institute of Statistical Research and Training, University of Dhaka, Dhaka; [2] Institute of Statistical Research and Training (ISRT), University of Dhaka, Dhaka

Introduction

Competing risks data arise when individuals are subject to multiple mutually exclusive event types, where the occurrence of one event precludes the observation of others. Standard approaches that treat competing events as independent censoring are well known to induce bias in cumulative incidence estimation. Random Survival Forests (RSF) offer a flexible, nonparametric framework for survival prediction however, existing extensions to competing risks settings often rely on simplifying assumptions, notably single imputation strategies and independence of censoring. This study aims to extend subdistribution-based RSF methodologies to better accommodate uncertainty in missing data and more realistic censoring mechanisms.

Methods

We propose two methodological enhancements to the competing risks RSF framework. First, a multiple imputation scheme is incorporated to appropriately propagate imputation uncertainty through the estimation of cumulative incidence functions (CIFs). Second, censoring is modeled as a function of covariates, thereby relaxing the conventional independent censoring assumption and allowing for informative censoring structures. The proposed approach is being evaluated through a comprehensive simulation design spanning a range of censoring intensities and dependency structures. Performance is assessed in terms of bias, variance, CIF estimation accuracy, and predictive discrimination relative to existing competing risks methods.

Results

The study is currently at an advanced methodological development stage, with the core framework, estimation procedures, and evaluation protocol fully specified. Preliminary implementation confirms the feasibility of integrating pooled CIF estimation across multiple imputations alongside covariate-adjusted censoring weights. The computational experiments indicate stable estimation behavior and suggest potential improvements in uncertainty quantification compared to conventional single-imputation approaches. Full-scale simulation results are in progress and are expected to provide a more definitive assessment of performance gains.

Discussion

This work directly addresses two critical limitations in current competing risks RSF methodologies: inadequate treatment of imputation uncertainty and the restrictive assumption of independent censoring. By jointly incorporating multiple imputation and covariate-dependent censoring, the proposed framework is positioned to yield more reliable and interpretable predictions in complex survival settings.

[30-P5-04] Analyzing length of stay with Ordinal Longitudinal Models: a comparison with competing risks in a simulation and real-data study

Presenter: Lorenzo Del Castello

Authors: Lorenzo Del Castello [1]; Davide Bernasconi [1]; Laura Antolini [1]; Paola Rebora [1, 2]

Affiliations: [1] School of Medicine and Surgery, Università Milano Bicocca, Monza, Italy., Monza; [2] Biostatistics and Clinical Epidemiology, Fondazione IRCCS San Gerardo dei Tintori, Monza, Lombardy, Italy., Monza

Introduction

Hospital length of stay (LOS) is a relevant indicator of healthcare quality, resource utilization, and treatment effectiveness. LOS is traditionally summarized using mean. However, this approach provides an incomplete picture and fails to account for in-hospital mortality. Competing risks account for discharge and death. In this framework, quantities of interest are the cause specific cumulative incidences and the survival functions, typically estimated with the Aalen-Johansen estimator. Ordinal longitudinal models (OLM) are an alternative promising approach by modelling ordered states (e.g. "Discharge" < "In hospital" < "Death") and estimating state occupation probabilities (SOPs). This study assesses the performance of OLM in modeling competing risks data through both a simulation study and an application to cardiology LOS data.

Methods

Simulations were conducted under two scenarios defining both Exponential and Weibull cause-specific proportional hazards, generating 1000 datasets, each composed of 1000 subjects, without censoring. Ordinal longitudinal models with time modeled using cubic splines were fitted. Model-derived SOPs at a specific time point were compared to the true values in terms of bias, mean squared error (MSE), and coverage, with 95% confidence intervals obtained with 1000 bootstrap resamples. A real data application on 455 patients with cardiogenic shock was performed. The ordinal outcome "Discharge Alive" < "CICU" < "Death" was modeled with OLM. Model-derived SOPs were compared with non-parametric Aalen-Johansen estimates.

Results

OLM performed well in estimating survival and cause specific cumulative incidences, with low biases and MSE. Coverage was satisfactory in the exponential scenario but suboptimal under the Weibull. In the real-data application, OLM-derived SOPs were coehrent with Aalen-Johansen estimates.

Discussion

The ordinal longitudinal model provides a flexible and effective approach for analyzing LOS data, offering a framework that can incorporate covariates in a unique model and can be extended to more complex multi-state settings.

References

Rohde, M. D., French, B., Stewart, T. G., & Harrell Jr, F. E. (2024). Bayesian transition models for ordinal longitudinal outcomes. Statistics in Medicine, 43(18), 3539-3561.

Harhay, M. O., Ratcliffe, S. J., Small, D. S., Suttner, L. H., Crowther, M. J., & Halpern, S. D. (2019). Measuring and analyzing length of stay in critical care trials. Medical care, 57(9), e53-e59.

[30-P5-05] Meta-analysis for competing risks: A how-to-guide for aggregate and individual participant data with examples from analyses comparing transcatheter and surgical treatment in patients with aortic-stenosis

Presenter: Matthias Klimek

Authors: Matthias Klimek [1]; Andreas Ziegler [1, 2, 3, 4]

Affiliations: [1] Cardio-CARE, Medizincampus Davos, Davos; [2] Department of Cardiology, University Heart & Vascular Center Hamburg, University Medical Center Hamburg- Eppendorf, Hamburg; [3] Center for Population Health Innovation (POINT), University Heart and Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg; [4] Discipline of Statistics, School of Agriculture and Science, University of KwaZulu-Natal, Pietermaritzburg

Introduction

Meta-analysis summarises the results of multiple independent clinical trials. In trials of survival outcomes, events may act as competing risks (e.g., stroke and death). Fixed effect (FE) and random effect (RE) meta-analyses are traditionally performed from aggregated data (AD) using published effect estimates. These estimates are generally unavailable for specific subgroups, and individual participant data (IPD) meta-analyses (IPDMA) need to be carried out. The main aim of this work is to provide a framework for 1-stage and 2-stage meta-analyses (MA) in the presence of competing risks for both FE and RE meta-analysis using IPD or AD.

Methods

For the RE analysis, we model correlated frailties for patients treated in the same trial or center, which represent unobserved covariates at the different cluster level (trial or center). This presentation builds on previous work by Ludwig et al. (2025 JAMA Cardiol). It provides competing risks estimators for subdistribution hazard ratios of the treatment effect of transcatheter vs surgical treatment of aortic-valve stenosis (TAVI vs SAVR) in 8 randomised controlled trials.

Results

Specifically, IPD from 4 investigator-initiated trials were meta-analysed using the 1-stage approach. AD from further 4 industrial sponsored trials were synthesized using classical MA, and an overall MA of all 8 trials was conducted using 2-stage IPDMA plus AD MA. Moreover, we discuss critically the recent use of reconstructed event rates from Kaplan-Meier survival curves, and we propose an approach for creating pooled cumulative incidence function curves across the trials.

Discussion

In summary, IPDMA is preferred over AD MA especially for subgroups, but both approaches can be integrated into a unified MA strategy for competing risks if estimates are available.

[30-P5-06] Accounting for Adverse Events in Survival Analysis: Comparing Two Methods in High-Risk Pediatric ALL

Presenter: Elena Tassistro

Authors: Elena Tassistro [1]; Maria Grazia Valsecchi [2, 3]; Laura Antolini [2]

Affiliations: [1] Fondazione IRCCS San Gerardo dei Tintori, Monza; [2] Bicocca Center of Bioinformatics, Biostatistics and Bioimaging (B4 centre), School of Medicine and Surgery, University of Milano-Bicocca, Monza; [3] Tettamanti Center, Fondazione IRCCS San Gerardo dei Tintori, Monza

Introduction

In clinical trials on novel treatments, outcomes commonly include a primary efficacy time-to-event endpoint (EP), related to disease control and survival, and a secondary endpoint related to occurrence of adverse events (AEs). For example, treatment-related infections in children with Acute Lymphoblastic Leukaemia (ALL) affect treatment outcomes by influencing quality and duration of survival. A common strategy to account for both types of events is to use a composite endpoint, combining EP and AE.

The standard analysis uses the time-to-first-event approach, considering the minimum time between EP and AE. It assumes equal importance of the events, which can artificially lower the survival curve due to occurrence of AEs. Alternatively, analysing the time-to-EP while ignoring AEs yields a more "optimistic" survival curve but neglects clinically important events.

Methods

Methods that account for different clinical relevance of individual components of a composite endpoint by assigning them different weights have been proposed: (i) weighted all-cause hazard, where pre-defined weights are multiplied by the cause-specific hazards [Rauch G et al. (2018). doi: 10.1002/sim.7531] (ii) weighted survival function, where non-parametrical estimates of the weighted survival probability are based on weighted event count and weighted number at risk [Bakal JA et al. (2015). doi: 10.1177/0962280211436004].

We reviewed these two approaches and developed a simulation protocol to evaluate the impact of the choice of weights on estimates obtained from (i) and (ii).

Results

Estimates obtained with (i) and (ii), for a given weight, are similar regardless of the value of the weight. In addition, they range between the estimates obtained from the time-to-first-event and the time-to-EP approaches. Results of the two approaches on a cohort of high-risk children with ALL, enrolled in the AIEOP-BFM ALL 2009 study in Italy, who experienced and primary and secondary endpoint events, are presented.

Discussion

Weighted approaches in composite endpoint analysis allow differentiation of the clinical importance of primary EP and AEs. Proper selection of weights is challenging but crucial since it significantly

influences survival probability and hazard estimates. Our simulation framework proved that both approaches work in a similar way in estimating survival curves.

[30-P5-07] Modelling multi-indication treatment in the presence of competing risks in multiple long-term conditions - a multistate modelling framework and patient-level simulation

Presenter: Rhiannon K Owen

Authors: Rhiannon K Owen [1]; Jane Lyons [1]; Ronan A Lyons [1]; Keith Abrams [2]

Affiliations: [1] Swansea University, Swansea; [2] University of Warwick, Coventry

Introduction

Healthcare decision-making often focuses on developing recommendations for single conditions, limiting the applicability in populations living with multiple long-term conditions (MLTCs). We address this by developing a multistate modelling framework to estimate intervention effects across joint disease trajectories in the presence of competing risks, motivated by the National Institute for Health and Care Excellence (NICE) guidance on sodium-glucose cotransporter-2 (SGLT2) inhibitors for type 2 diabetes mellitus (T2DM), chronic kidney disease (CKD), and heart failure (HF).

Methods

Markov multistate models adjusted for age, sex, and deprivation were used to estimate the transition between T2DM, CKD, HF, their combinations, and death. Transition specific baseline hazards were estimated using parametric survival models applied to population-scale, individual-level, linked electronic health records (EHR) for 1.3 million adults in Wales over a 10-year period (2002 to 2012). Treatment effects were incorporated as transition-specific hazard ratios, assuming proportional hazards, derived from published randomized controlled trials in combinations of MLTCs. A patient-level simulation was used to propagate these effects to estimate marginal outcomes under competing risks for treatment versus no treatment regimens. Estimands included life expectancy, time to next condition, and time spent in each disease state.

Results

SGLT2 inhibitors decreased the estimated mean life expectancy by -0.20 years (95% Confidence Interval (CI): -0.23, -0.17) in people living with T2DM, CKD, and HF, and their combinations. Fewer people developed all three conditions on SGLT2 inhibitors compared to those untreated, with an increase in the estimated mean time to develop the next condition in the sequence by up to 0.72 years (95% CI: 0.64, 0.80), which was observed for the development of HF in those with CKD. The estimated mean time to develop T2DM as the third condition in this sequence increased by 0.55 years (95%CI: 0.52, 0.58) for those treated with SGLT2 inhibitors compared to untreated.

Discussion

This framework integrates multistate disease progression with externally derived treatment effects to estimate holistic policy-relevant outcomes in MLTC populations under competing risks. Whilst

SGLT2 inhibitors resulted in a minimal reduction in estimated mean life expectancy, time spent in health states with increased quality of life (such as those with fewer diseases) increased.

[30-P5-08] Analysing trajectories of multiple cancer diseases using a landmark multi-state modelling approach: An example analysis of SEER registry data

Presenter: Leonas Lanwer

Authors: Leonas Lanwer [1]; Dennis Görlich [2]; Andreas Faldum [3]

Affiliations: [1] University of Münster, Münster; [2] Universität Münster, Münster; [3] Institut für Biometrie und Klinische Forschung (IBKF), Universität Münster, Münster

Introduction

The analysis of patient data with information on transitions between health states can be addressed by multi-state models. Newer methodological advancements, like the landmark Aalen-Johansen (LMAJ) estimator (Putter & Spitoni 2018.DOI: 10.117/0962280216674497), aim to loosen the assumption of the Markov property. Our aim was to estimate transition probabilities using the LMAJ estimator in a multi-state model based on real-world registry data for multiple cancer diseases after breast cancer.

Methods

We accessed patient data from the US Surveillance, Epidemiology, and End Results Program (SEER, Research Data, 8 Registries, Nov 2023 Sub). A structural "breast cancer model" for a maximum of 3 sequential cancers, representing the most abundant transitions after a breast cancer diagnosis, was developed. Rare transitions were gathered in summary states. Two absorbing states, representing death attributed to any cancerous or non-cancerous cause, were included in the model. The Markov assumption was tested transition-wise with log-rank tests (Titman & Putter 2022.DOI:10.1093/biostatistics/kxaa030). Model estimates were computed using the R-package mstate (de Wreede et. al. 2011.DOI:10.18637/jss.v038.i07). Results were visualized as annotated network-diagrams, landmark-specific transition probability plots over time and contour plots of time-dependent transition probabilities and landmark times.

Results

In total 586,688 patients were included in the breast cancer model, 82,588 patients developed at least a second cancer. 29 of the model transitions violated the Markov assumption. The largest probabilities were found for transitions from non-absorbing into absorbing states for selected landmarks. Contour plots showed distinct patterns of transition probabilities in dependence of the landmark time and follow-up time. These patterns indicate the existence of distinct time intervals defining areas-of-interest with higher chances of transition.

Discussion

Our analysis enabled the model-specific assessment of the probabilities of subsequent cancers and absorption by two categories of death-causes. Additionally, the identification of transition-specific higher-risk intervals, defined by the landmark time and the observation time, was made possible. As limitations, some assumptions of the LMAJ estimator may not fully hold in our model (e.g., random right-censoring). Furthermore, the landmarked analysis can reduce precision and introduce potential bias. We believe that our approach could also be used in other non-strictly Markovian data, where many possible subject trajectories are of interest.

[30-P5-09] Integrated Modeling of Functional Trajectories in Palliative Care: A Multi-State Model Analysis of Ambulation, Nutrition, and Communication

Presenter: Katsuei Takahashi

Authors: Katsuei Takahashi [1]; Ayano Takeuchi [1]; Jun Hamano [2]

Affiliations: [1] Chuo University, Graduate School of Science and Engineering, Tokyo; [2] University of Tsukuba Hospital, Ibaraki

Introduction

Understanding prognosis is essential in palliative care, particularly regarding "functional prognosis," which describes the decline in physical functions over time. Unlike overall survival, which uses death as an endpoint, functional prognosis represents the time-to-event for the loss of essential physical and cognitive functions, such as ambulation, oral intake, and communication. From the perspective of advance care planning (ACP), accurate prognostic information is crucial to help patients and families anticipate the dying process and make appropriate decisions. This study aimed to develop an integrated prognostic model that simultaneously estimates three interrelated functional outcomes - ambulation, eating, and communication - based on baseline clinical information in palliative care patients.

Methods

Data were obtained from the EASED (East Asian Collaborative Study to Elucidate the Dying Process), a multicenter prospective observational study conducted in 2017. The original cohort included 1,896 terminally ill cancer patients admitted to 23 palliative care units in Japan. Among these, 125 patients with complete data were included in the analysis. A multi-state model was employed to characterize the stochastic process of functional decline, where transitions between states (functional loss and death) were formulated as transition intensities.. This approach allows simultaneous analysis of multiple interdependent functional declines as a stochastic process. Two models were constructed: a baseline model without covariates and a model including metastases (liver, bone, lung, and nervous system) as covariates.

Results

Model fit was assessed via log-likelihood the inclusion of metastatic sites as covariates yielded a statistically superior fit (Log-likelihood: -1015.54 vs. -1025.66), suggesting that metastatic patterns influence functional transition rates. Notably, bone metastasis had a significant effect: among patients who had lost ambulation and eating function, the hazard for subsequent loss of communication was significantly elevated (HR 2.11 95% CI: 1.01-4.41, p = 0.046).

Discussion

This indicates that bone metastasis may act as a catalyst for the rapid deterioration of remaining functions in the final stages of life, potentially due to skeletal-related events and associated pain. These findings highlight the potential of the proposed model to support understanding of functional trajectories and enhance ACP in palliative care.

[30-P5-10] Potential of (Joint) Frailty Models to Elucidate Sex-Specific Survival Differences in Bladder Cancer

Presenter: Ian Wittenberg

Authors: Ian Wittenberg [1]; Ahmed Bedir [1]; Alexander Kluttig [1]; Maria Elena Lacruz [1]; Saskia Thies [1]; Andreas Wienke [2]

Affiliations: [1] Cancer Registry Saxony-Anhalt, Magdeburg; [2] Martin-Luther University Halle-Wittenberg, Institute of Medical Epidemiology, Biostatistics and Informatics, Halle (Saale)

Introduction

Urothelial bladder cancer (UBC) is marked by frequent multifocality and recurrence, with a higher incidence observed among males compared to females. Factors causing sex-specific survival differences have been discussed, including diagnostic delay, biological factors, and treatment patterns, but the disparities remain insufficiently understood [1]. Although recurrent events are central in UBC, frailty models (FM) and joint frailty models (JFM) are rarely used in cancer epidemiology. These approaches allow simultaneous analysis of recurrence and overall survival (OS) and may help to clarify whether survival differences are related to recurrences, treatment patterns or unobserved heterogeneity.

Methods

Cancer registry data from Saxony-Anhalt included 1,183 women and 3,076 men diagnosed with UBC (ICD-10 C67, D09.0) between 2018 - 2023 sarcomas and adenocarcinomas were excluded. During follow-up until November 2025, 509 women and 1,169 men died. We conducted preliminary analyses using Cox proportional hazards models and FM to estimate sex-specific OS. Models were adjusted for UICC stage at diagnosis, histologic grade, and age. Effect modification by sex and stage was assessed through interaction terms. FM were specified at the patient level to account for unobserved heterogeneity. Final analyses will apply JFM to jointly model recurrent events and death and incorporate treatment information.

Results

In univariable Cox models, women had lower OS than men (HR 1.22, 95% CI 1.10-1.35). Adjustment for stage, histologic grade and age reduced the sex-specific effect on OS (HR 0.97, 95% CI 0.87-1.08). Adding a frailty term increased the univariable HR to 1.78 (95% CI 1.43-2.20), but not the adjusted HR (0.98, 95% CI 0.85-1.13). Interaction models showed consistent sex-specific differences across stages I-IV (HRs 1.07-1.16). Analyses of recurrence rates and treatment patterns are ongoing.

Discussion

The results indicate that sex-specific survival differences vary by stage at diagnosis. Given the recurrent nature of UBC and potential stage changes over time, JFM may clarify whether sex differences arise from recurrences, postrecurrence treatments, or unobserved heterogeneity. Identifying the mechanisms behind sex-specific OS differences may help reduce this survival gap.

[1] Ceria F et al. Unveiling the Female Factor: Gender-Based Differences in Outcomes and Survival Following Radical Cystectomy for Bladder Cancer. Cancers (Basel). 2026 Jan 2018(2):308. doi: 10.3390/cancers18020308.

[30-P5-11] Repeated histological diagnoses and kidney graft failure: an observational cohort study

Presenter: Arthur Vranken

Authors: Arthur Vranken [1]; Maarten Coemans [1]; Priyanka Koshy [2, 1]; Angelica Pagliazzi [1]; Thomas Vanhoutte [1, 2]; Karolien Wellekens [1, 2]; Thibaut Vaulet [1]; Dirk Kuypers [1, 2]; Frederike J. Bemelman [3]; Bertrand Chauveau [4]; Christophe Masset [5]; Tim Debyser [1]; Sandrine Florquin [3]; Jesper Kers [6, 3]; Aiko P.J. de Vries [6]; Soufian Meziyerh [6]; Geert Verbeke [1]; Maarten Naesens [1, 2]

Affiliations: [1] KU Leuven, Leuven; [2] University Hospitals Leuven, Leuven; [3] Amsterdam University medical centers, Amsterdam; [4] CHU de Bordeaux, Bordeaux; [5] CHU de Nantes, Nantes; [6] Leiden University Medical Center, Leiden

Introduction

In the existing literature, the analysis of histological diagnoses in sequential kidney transplant biopsies often involves methods that reduce power or induce bias (landmarked survival models, ignoring correlated observations within patients, immortal time bias, discretization of the time scale). Survival models with time-dependent covariates and recurrent events provide an elegant alternative to the analysis of sequential biopsies.

Methods

In two observational cohorts (1819 transplantations with 6272 post-transplant biopsies and 751 transplantations with 851 biopsies), Cox conditional frailty models with time-dependent covariates quantified associations of prior diagnoses of T-cell mediated rejection (TCMR), Borderline TCMR (bTCMR), antibody-mediated rejection (AMR), Probable AMR (pAMR), DSA-negative C4d-negative microvascular inflammation (MVIDSA-/C4d-), and BK polyomavirus nephropathy (BKPyVAN) with hazards of

subsequent diagnoses. In addition, Cox cause-specific hazard models with time-dependent covariates estimated associations of sequences of diagnoses with graft failure.

Results

The hazard of a diagnosis was most strongly associated with earlier diagnosis of the same type, but some diagnoses were also affected by prior diagnoses of a different type. The hazard of kidney graft failure was significantly increased by a first and second diagnosis of TCMR, AMR or MVIDSA-/C4d- in both cohorts. In these analyses, the association was significantly stronger after a second diagnosis than after a first diagnosis. In the protocol biopsy-rich Leuven cohort, associations persisted for diagnoses observed in protocol biopsies, in the absence of clinical symptoms.

Discussion

Incidence of Banff histological diagnoses appears to be affected by earlier diagnoses, especially those of the same type. Occurrence of multiple sequential diagnoses was strongly associated with an increased hazard of graft failure, more so than only a single diagnosis. This underscores shortcomings of current treatment options for these recurrent or persistent injury processes.

[30-P5-12] Effect of time-dependent concomitant cardiovascular medication on time-to-first-event endpoints in the Cox regression model based on the example of DIGIT-HF

Presenter: Annabell Wede

Authors: Annabell Wede [1, 2]; Cornelia Frömke [2]; Udo Bavendiek [3]; Dominik Berliner [3]; Johann Bauersachs [3]; Armin Koch [1]; Anika Großhennig [1]; Nele Henrike Thomas [1]

Affiliations: [1] Institute for Biostatistics, Hannover Medical School, Hannover; [2] Hochschule Hannover - University of Applied Sciences and Arts, Hannover; [3] Department of Cardiology and Angiology, Hannover Medical School, Hannover

Introduction

The double-blind, randomized, controlled, multicenter DIGIT-HF trial recently demonstrated that the cardiac glycoside digitoxin reduced the primary composite outcome of all-cause death (ACD) or first hospitalization for worsening heart failure (HFH) compared to placebo in patients with heart failure (HF) and a reduced ejection fraction receiving a guideline-recommended HF therapy. This concomitant HF medication was intended to be optimized during the study period in accordance with current guidelines.1 Due to the long study duration, it can be assumed that the HF medication changed over time at patient level. Thus, the longer the survival time of patients, the higher the chance of receiving an optimized HF medication or of a fading optimized HF medication due to comorbidities or compliance. To correctly model the influence of such a time-dependent covariate, the Cox regression model used for the primary analysis of DIGIT-HF has to be extended accordingly.2

Methods

After descriptively analyzing the HF medication and its potential change over time, the Cox regression model of the primary analysis will be adjusted for the time-dependent covariates triple and quadruple HF therapy using the counting process approach2 (PROC PHREG procedure in SAS®). Outcomes of interest are the primary endpoint ACD or first HFH and the key secondary endpoint ACD, with focus on the triple and quadruple therapy as they cover clinically reasonable combinations of HF medication. Triple therapy is defined as ACE-inhibitor or AT-1-receptor-blocker or ARNI and beta-blocker and MRA. Quadruple therapy is specified as ACE-inhibitor or AT-1-receptor-blocker or ARNI and beta-blocker and MRA and SGLT2-inhibitor.1

Results

The conduct of adjusting the Cox regression model for the time-dependent covariates triple and quadruple therapy for the primary endpoint ACD or first HFH and the key secondary endpoint ACD is currently ongoing. The results will be presented by reporting the number of patients, events per 100 patient-years by treatment (digitoxin/placebo) and the hazard ratio with the 95% confidence interval.

Discussion

The results will be compared and discussed with the results of DIGIT-HF and other trials investigating HF patients, specifically those where time-dependent covariates were analyzed.

1. Bavendiek, U. et al. doi: 10.1056/NEJMoa2415471.

2. Kleinbaum, D. G. & Klein, M. doi:10.1007/978-1-4419-6646-9.

[30-P5-13] Time-dependent effects of immunosuppressive therapies on primary incidence of cancerous disorders and death following liver transplantation

Presenter: Zdeněk Valenta

Authors: Tereza Hakova [1, 2, 3]; Pavel Taimr [1]; Jiří Froněk [4]; Zdeněk Valenta [5]

Affiliations: [1] Dept. of Hepatology and Gastroenterology, Institute of Clinical and Experimental Medicine (IKEM), Prague; [2] First Faculty of Medicine, Charles University, Prague; [3] Erasmus MC, Department of Gastroenterology and Hepatology, Rotterdam; [4] Dept. of Transplant Surgery, Institute of Clinical and Experimental Medicine (IKEM), Prague; [5] Dept. of Statistical Modelling, Institute of Computer Science of the Czech Academy of Sciences, Prague

Introduction

De novo malignancies (DNM) represent a major complication following liver transplantation (Tx), with immunosuppressives playing a significant role in their development. While immunosuppression is essential for preventing graft rejection, cumulative exposure may increase the risk of malignancy and mortality. This study evaluates the competing risks of primary incidence of de novo malignancy and death in relation to cumulative immunosuppressive exposure.

Methods

We analyzed left-truncated and right-censored data from a retrospective cohort of 1,034 adult liver transplant recipients. Follow-up was restricted to 10 years post-TX. The primary endpoint was the first occurrence of DNM. Cumulative exposure to immunosuppressive agents was modeled using time-dependent covariates within both multiplicative and additive survival frameworks. Analyses were adjusted for relevant confounders and employed cause-specific Cox proportional hazards models, as well as Aalen's and McKeague-Sasieni additive models to characterize time-varying effects.

Results

Recipient age and male gender were associated with significantly elevated, time-invariant hazard of DNM incidence. In contrast, cumulative exposure to Everolimus and Mycophenolate was protective. Mycophenolate showed a neutral or slightly adverse association early after TX, becoming protective after approximately 4-5 years. For the competing event of death, older recipient age and active smoking were associated with hazards increasing over time, reaching statistical significance after approximately 4.5 and 6 years, respectively. Everolimus, Mycophenolate, and Tacrolimus were each associated with reduced hazard of death. Although all three agents demonstrated time-varying protective effects, these were most pronounced early after TX. Everolimus and Mycophenolate maintained protective associations throughout the follow-up, with modest attenuation over time. Tacrolimus conferred protection for approximately 6 years, after which period its effect diminished.

Discussion

The effects of immunosuppressive therapies on post-transplant malignancy and mortality were drug-specific and time-dependent. Everolimus and Mycophenolate consistently demonstrated protective associations, whereas recipient-related factors - age, male gender, and smoking - were associated with adverse outcomes. These findings highlight the value of flexible survival modeling and support individualized, time-adapted immunosuppressive strategies. Additive survival models corroborated the results and provided additional insight into temporal dynamics. Notably, no immunosuppressive agent exhibited adverse effects on either outcome.

[30-P5-14] Mortality and causes of death in idiopathic inflammatory myopathies: patients' survival and causes of death from the French national health data system linked to the French national registry for rare diseases

Presenter: Anne-Sophie Jannot

Authors: Dylan Vellas [1]; Cyril Gitiaux [2]; Ludivine Kouton [3]; Olivier Benveniste [4]; Yves Allenbach [4]; Anne-Sophie Jannot [5]

Affiliations: [1] Inserm, Université Paris Cité, Inria, UMR 1346, HeKA, Paris; [2] Reference Center for neuromuscular diseases, Necker-Enfants-Malades University Hospital, APHP, FILNEMUS, Paris, France Department of Pediatric Neurophysiology, Necker-Enfants malades Hospital, University of Paris, AP-HP, Paris, France., Paris; [3] Referral centre for neuromuscular diseases and ALS, La Timone hospital, Aix-Marseille University, ERN-NMD, FILNEMUS, Marseille, France, Marseille; [4] UMR974, Sorbonne University, AP-HP, Paris, France Department of Internal Medicine and Clinical Immunology, National Reference Centre for Inflammatory Myopathies, hôpital Pitié-Salpêtrière, Paris, France, Paris; [5] HeKa Université Paris Cité, INRIA, INSERM, F-75015, Paris

Introduction

National registries such as the French National Registry for Rare Diseases (BNDMR) enable precise patient identification, while healthcare claim databases like the French National Health Data System (SNDS) provide longitudinal care trajectories. Their linkage offers valuable insights into rare diseases mortality. However, methodological challenges include immortal time bias when diagnoses predate registry implementation and missing clinical data such as diagnosis date.

This study aims to estimate the survival in myositis compared with the general population and to identify their causes of death using nationwide data while tackling these two challenges.

Methods

This work is part from the Drómos study that linked the French National Registry for Rare Diseases (BNDMR) with the National Health Data System (SNDS).

Subgroups were identified by Orpha codes from BNDMR data (2010-2022). Cases were matched to controls on birth year, sex and city of birth.

Survival was estimated from IIM diagnosis and compared by Kaplan-Meier estimators and a Cox model adjusted for comorbidities estimated by hospitalisation and chronic-disease status. Follow-up was considered from the first consultation in expert centre to avoid immortal time bias.

Cumulative Incidence Functions and univariate Fine & Gray model compared causes of death. Causes of death ratios were estimated. The Standardized Ratio of Multiple to Underlying (SRMU) assessed the role of each recorded cause of death between causal or participating cause.

Results

There were 1,888 myositis cases and 5,538 controls included (diagnosis age 58.9y [37.3580.55] female predominance).

Survival was 96% one year after diagnosis (95% CI[94.4%97.6%]), 85.1% five years after (95% CI[82.6%87.7%]).

Myositis survival was decreased compared to controls (HR 3.35[2.754.08]).

DM displayed the higher Hazard Ratio (HR 5.15[3.138.49]) followed by IMNM (HR 4.27[2.357.77]), ASyS (HR 3.74[2.345.96]) and IBM (HR 2.68[2.043.52]), non-modified by adjustment.

Respiratory causes were the most specific cause of death (sHR 5.36[2.6410.9]) followed by cancer (sHR 3.59[2.275.7]).

SRMU for cancer was 2.16[1.473.17] and 4.10[2.516.68] for respiratory causes.

Discussion

This nationwide study showed an increased mortality in every subgroup. For the first time, an increased mortality in ASyS was identified as well as in IMMNs and IBMs. Respiratory causes were the most specific causes of death, often participating in the death process.

[30-P5-15] Time to cure in adolescents and young adults (AYA) in Germany

Presenter: Nora Eisemann

Authors: Nora Eisemann [1]; Hannah Baltus [1]; Antje Schliemann [1]; Laura Schumann [1]; Paula Grieger [1]; Annika Waldmann [1]; Ron Pritzkuleit [2]; Inken Hilgendorf [3, 4]; Alexander Katalinic [1, 2]

Affiliations: [1] Institut für Sozialmedizin und Epidemiologie, Universität zu Lübeck, Lübeck; [2] Institut für Krebsepidemiologie e. V., Universität zu Lübeck, Lübeck; [3] Deutsche Stiftung für junge Erwachsene mit Krebs, Berlin; [4] Universitätsklinikum Jena, Jena

Introduction

Adolescents and young adults (AYA) who have survived cancer may face rejection by insurance or credit companies because of a perceived higher risk of morbidity or death. While it is impossible to determine cure in an individual cancer survivor, "statistical cure" can be estimated.

The aim of this analysis is to provide evidence on the estimated "time to cure" for the most common cancer groups of AYA in Germany.

Methods

AYA are defined as patients aged 15 to 39 years at diagnosis. "Statistical cure from cancer" is defined as having the same mortality probability as a sex- and age-matched individual from the general population.

Cancer registry data from all federal states in Germany over a 20-year-period (2003-2022) were provided by the Center of cancer registry data.

In addition to non-parametric net survival estimates (Pohar-Perme) by calendar year, we currently apply cure regression models to estimate the time to cure for the most common cancer groups.

Results

Nearly 310,000 AYA were included, 58.3 % of which were female. With higher age at diagnosis, the number of patients increased exponentially (from 2420 aged 15 years to 30,633 aged 39 years). Preliminary analyses of Pohar-Perme estimates of the net survival indicate a very small excess hazard of death for testis cancer (germ cell and trophoblastic of testis) and thyroid cancer, a time to cure of approximately 10-15 years for melanoma, cervical and ovary cancer (germ cell and trophoblastic of ovary), and longer times to cure for breast cancer, colorectal cancer, leukemia, soft tissue sarcoma and bone sarcoma. Final results will be presented at the conference.

Discussion

A pre-condition of using cure models is the manifestation of a plateau in net survival over time. Despite a maximum follow-up of 20 years, such a plateau was not for all cancer groups clearly reached. When net survival approaches a plateau very slowly, the time to cure estimate comes with a large degree of uncertainty.

[30-P5-16] Cancer survival: Comparison of continuous vs. categorical modelling of cancer onset year in the SEER registry data

Presenter: Dennis Görlich

Authors: Dennis Görlich [1]; Leonas Lanwer [1]; Karen Fischhuber [1]; Eike Bormann [1]

Affiliations: [1] University of Münster, Münster

Introduction

Mortality improvement over the last decades has been described in the literature for all cancer-entities. Recently, we reported an analysis of SEER data on acute myeloid leukemia using the year of diagnosis as categorical factor (Görlich et al., 2025. DOI:10.3390/cancers17233735). With the present analysis, our aim was to assess the chosen methodology and to compare the yearly categorical modelling approach with a continuous approach.

Methods

We analyzed the US SEER data set, including data of more than 4 Mio. cancer patients diagnosed between 1975 and 2021. Overall survival was analyzed in a Cox-model to estimate hazard ratios (HRcat(y)=frach(y)h(1975)) for (categorical) year of diagnosis with a common reference year 1975. HR from the continuous approach (HR=frach(y+1)h(y)) were projected onto the categorical x-axis by computing HRp(y)=HR(y-1975). Differences between HRcat(y) and HRp(y)were evaluated. Analyses were repeated for selected cancer types.

Results

For testicular cancer (N=30,014), we found a HR of 0.98 (95%CI: 0.977-0.983) per diagnosis year. The corresponding HRcat(y) showed an early decline (until the late 1980's), likely due to approval of cisplatin, and a following plateau. The average HRcat(y) on the plateau was 0.42(SD:0.05). The RMSD between HRp(y)and HRcat(y) was 0.30 with a mean absolute deviation of -0.27. In contrast, continuous and categorical models for breast cancer showed very comparable trajectories (N=725,410 RMSD: 0.07 mean abs. dev.: 0.04). Other cancer types will be reported and discussed as well as the systematic evaluation of differences between categorical and continuous modelling.

Discussion

Our analysis showed that HR and HRcat represent different aspects of survival and the impact of medical progress. While the former gives an average improvement rate per year, the latter summarized the absolute progress over a specific period of time. The comparison of the assumed homogenous process (continuous model) and the flexible estimates of the HRcat allowed us to quantify the deviance between the two assumed process types. The difference between categorical and continuous estimates can be used as an indicator of potentially non-homogenous processes of survival improvement.

[30-P5-17] Breast cancer survival in Ireland: a population-based analysis using National Cancer Registry Ireland data

Presenter: Arya C Soman

Authors: Arya C Soman [1]; Shirin Moghaddam [1]; Amirhossein Jalali [1]; Aedin Culhane [1]

Affiliations: [1] University of Limerick, Limerick

Introduction

Breast cancer is the most commonly diagnosed cancer among women in Ireland. Although improvements in screening, diagnostic pathways, and treatment have contributed to changes in incidence, mortality, and survival over the past two decades, inequalities in outcomes persist, particularly across socioeconomic groups. Understanding how these trends have evolved, and how Ireland compares with other OECD countries, is essential for informing equitable cancer control policies.

Methods

This study undertakes a population-based retrospective analysis using data from the National Cancer Registry Ireland (NCRI), supplemented by OECD indicators. Survival outcomes will be analysed using Kaplan-Meier estimation and log-rank tests for initial comparisons, followed by multivariable Cox proportional hazards models to estimate overall survival. Competing risks regression (Fine-Gray) will be used to model breast cancer-specific mortality, and mean residual life (MRL) functions will be applied to provide clinically interpretable measures of remaining survival time.

Models will adjust for demographic, clinical, and tumour characteristics (including grade, stage, lymph node involvement, and ER/PR/HER2 status), as well as route of diagnosis and treatment modalities. Socioeconomic deprivation will be examined as a key determinant of survival inequalities, and temporal trends will be analysed to assess changes over time, with international benchmarking using selected OECD indicators.

In addition, predictive modelling approaches, including regularised regression (Lasso Cox), random survival forests, and gradient boosting methods, will be employed to identify determinants of early-stage detection and survival outcomes.

Results

Data preparation and model development are ongoing. Preliminary exploration indicates variability in stage at diagnosis, presentation pathways, treatment patterns, and socioeconomic distribution. The proposed modelling framework will quantify the relative contributions of screening, stage, treatment, and socioeconomic factors to survival outcomes.

Discussion

This study presents a comprehensive population-based modelling framework integrating survival analysis and predictive modelling to evaluate breast cancer outcomes. By combining classical survival models with mean residual life approaches, the framework provides a clinically interpretable understanding of survival beyond traditional measures.

The analysis explicitly addresses socioeconomic inequalities and variations in diagnostic and treatment pathways, offering insights relevant to screening policy, treatment optimisation, and equitable cancer care. This work contributes to advancing real-world evidence generation using robust statistical methodologies in population health research.

[30-P5-18] Robustness of stage as a prognostic factor for lung cancer: a multiverse survival analysis based on SEER data

Presenter: Anna Sordo

Authors: Anna Sordo [1, 2]; Stefania Lando [3, 2, 4, 5]; Mariavittoria Vescovo [3, 4]; Jiregna Olani Kedida [3, 4]; Ester Rosa [3, 2, 4]; Gloria Brigiari [3, 2, 4]; Dario Gregori [3, 2]

Affiliations: [1] Unit of Biostatistics, Epidemiology and Public Health, Department of Cardiac, Thoracic, Vascular Sciences and Public Health, University of Padova, Padua; [2] BIOSTAT-X Biostatistics & AI for Biomedical Discovery, Institute of Pediatric Research (IRP) 'Città della Speranza', Padova, Italy, Padova; [3] Unit of Biostatistics, Epidemiology and Public Health, Department of Cardiac, Thoracic, Vascular Sciences and Public Health, University of Padova, Padova; [4] PhD Program in Translational Specialistic Medicine "G.B. Morgagni", Curriculum "Biostatistics and Clinical Epidemiology", University of Padova, Padova; [5] Department of Surgery, Oncology and Gastroenterology (DiSCOG), University of Padova, Padova, Italy, Padova

Introduction

Tumour staging is a key prognostic factor for survival, but it is heterogeneous and not collected in a uniform way worldwide. Missing or incomplete information used to determine staging may limit a reliable assessment of its stability as a prognostic factor. This study evaluates the robustness of stage as a prognostic factor for survival across a set of plausible analytical scenarios.

Methods

A cohort of 99,985 lung cancer patients diagnosed between 2010 and 2015 was selected from the SEER (Surveillance, Epidemiology, and End Results) database. Only patients with complete data on tumour stage and covariates required for the different model specifications were included. Tumour

stage was classified according to the 7th edition of the AJCC/UICC TNM staging system. A multiverse analysis was applied with Cox proportional hazards models to evaluate the prognostic value of stage on overall survival, exploring multiple analytical specifications based on different sets of demographic and clinical/histological covariates. The prognostic effect of stage was assessed through three pairwise comparisons (stage I vs II, stage II vs III, and stage III vs IV).

Results

Higher tumour stage was consistently associated with worse overall survival across all analytical specifications, with hazard ratios (95% CI) for stage I vs II (HR = 1.49, 95% CI 1.39-1.60, $p < 0.01$), stage II vs III (HR = 1.68, 95% CI 1.55-1.83, $p < 0.01$), and stage III vs IV (HR = 2.21, 95% CI 2.15-2.29, $p < 0.01$). The magnitude and direction of these estimates remained stable across models, supporting the robustness of stage as a prognostic factor.

Discussion

These findings support multiverse analysis as a useful approach to investigate the robustness of staging variables in observational research. Tumour stage emerged as a robust prognostic factor across multiple plausible model specifications, suggesting limited sensitivity to common analytical choices. Even in the presence of missing or incomplete data, including multiple covariates allowed the observed associations to remain broadly consistent.

[30-P5-19] Extended Joint Nested Mixed-Effects (JoiNME) model for high-dimensional longitudinal biomarkers and time-to-event data using a shared latent process

Presenter: Trinh Dong

Authors: Trinh Dong [1, 2, 3]; Joseph Donovan [3]; Liane Canas [4]; Thuong-Thuong Nguyen [3]; Marc Modat [4]; Guy Thwaites [5, 3]; Ronald Geskus [5, 3]

Affiliations: [1] Population Health Sciences Institute - Biostatistics Research Group, Newcastle University, Newcastle upon Tyne; [2] Univ. Bordeaux, INSERM, Institut Bergonié, BPH, U1219, CIC 1401, Bordeaux; [3] Oxford University Clinical Research Unit, Ho Chi Minh City; [4] King's College London, London; [5] University of Oxford, Oxford

Introduction

Advances in medical technology increase the interest in investigating high-dimensional longitudinal biomarkers. For instance, recent clinical trials of tuberculous meningitis collected serial brain magnetic resonance images to compare the longitudinal trajectories of 60 regional brain volumes between treatment arms. Using standard joint models (JMs) [1], fiting K correlated biomarkers involves a K x K correlation matrix. A large K can inflate computation and reduce statistical power.

To address this, we propose an extended JM where longitudinal components are represented by a two-level nested mixed-effects model (JoiNME). Observed measurements are assumed following a common latent trajectory, combined with marker-specific heterogeneity, participant-specific heterogeneity, and participant-dependent dispersion.

Methods

Participant survival is fitted using Cox model. We formulate the longitudinal sub-model with three compartments (i) a population-level mean trajectory shared across participant i and biomarkers k - for simplicity is denoted as eta(t) (ii) participant-specific (fi(t)) and marker-specific (gk(t)) random effects and (iii) participant-marker interaction random effects hik(t). Variance-covariance matrix (Sigmai) for hik(t) can vary per participant using secondary mixed models.

Association between longitudinal and survival processes is parametrised via: (a) current value and/or slope of participant-level mean trajectory across markers (eta(t)+fi(t)) (b) current value and/or slope of participant-marker random effects (hik(t)) with extensions to (c) components of participant-marker correlation matrix (Ki) and (d) components of Sigmai. Non-linearity is handled by functional transformation [2] or monotonic splines. Cumulative hazard is integrated using Gauss-Kronrod quadrature method [1]. Bayesian estimation by Hamiltonian Monte Carlo (HMC) is performed using Stan and packaged into the R library joinme.

Results

In simulations with 60 biomarkers and linear association, JoiNME showed better convergence than JMbayes2 [2] and maintained smaller bias and coverage rate, especially in the association parameters.

Discussion

Leveraging common latent process with nested random effects, JoiNME provides an efficient and flexible method to jointly model high-dimensional biomarkers and survival outcome, avoiding fitting a large correlation matrix, and improving statistical power. Limitations include non-guaranteed identifiability, shrinkage-induced bias, and slow HMC adaptation.

[1] Rizopoulos, D. (2012). Joint Models for Longitudinal and Time-to-Event Data: With Applications in R (1st ed.). Chapman and Hall/CRC.

[2] Rizopoulos D, Miranda-Afonso P, Papageorgiou G (2026). JMbayes2: Extended Joint Models for Longitudinal and Time-to-Event Data.

P6 - Evidence and Decision Sciences

Poster area: P6

[30-P6-01] Targeted test evaluation: Five suggestions for refining the framework for designing diagnostic accuracy studies with clear study hypotheses

Presenter: Werner Vach

Authors: Werner Vach [1, 2]

Affiliations: [1] Basel Academy for Quality and Research in Medicine, Basel; [2] aDepartment of Sport Sciences and Clinical Biomechanics, University of Southern Denmark, Odense

Introduction

The design of single-arm diagnostic accuracy studies is challenging. Korevaar and colleagues [1] proposed a framework for designing diagnostic accuracy studies, focusing on the definition of clear study hypotheses. Such clear study hypotheses allow to base the final analysis of a diagnostic accuracy study on formal hypothesis testing and consequently to perform sample size calculations. They proposed to derive minimally acceptable criteria for sensitivity and specificity to specify the hypothesis of interest and described the process to arrive at such values. The basic cornerstone is a weighing of the potential consequences of false positive and false negative test decisions, which should be based on a good understanding of the target condition of the test, the target patients, the role of the test, and downstream consequences of test results.

Methods

Recently, five potential refinements have been proposed [2]. They aim at increasing the flexibility of the framework while pertaining or improving its logical consistency:

1) The relationship between minimal criteria and the choice of the null hypothesis region can be made more explicit by defining a region of clinical uselessness and a region of clinical usefulness.

2) A potential compensation between sensitivity and specificit can be allowed by adjusting the shape of the region of clinical uselessness.

3) If weighing between false postive and false negative decisons in based on other parameters than sensitivity and specificity, they can be directly used to define the regions.

4) The analytical strategy of the study can be phrased as an estimation problem without any change in sample size considerations.

5) Directly moving to a comparative accuracy study may facilitate the phrasing of study hypotheses and minimal criteria.

Results

The poster illustrates some of the potential refinements.

Discussion

The design of diagnostic accuracy studies can benefit from the five proposed refinements.

References:

1. Korevaar DA, Gopalakrishna G, Cohen JF, Bossuyt PM. Targeted test evaluation: a framework for designing diagnostic accuracy studies with clear study hypotheses. Diagn Progn Res 3, 22 (2019).

2. Vach, W. Targeted test evaluation: five suggestions for refining the framework for designing diagnostic accuracy studies with clear study hypotheses. Diagn Progn Res 9, 27 (2025).

[30-P6-02] Statistical methods for the analysis of diagnostic test accuracy in clustered data settings - A systematic review of methods

Presenter: Daniel Dümmler

Authors: Daniel Dümmler [1]; Philipp Weber [2]; Frederike Vogel [2]; Valeriia Dobler [2]; Antonia Zapf [2]; Nicole Rübsamen [1]

Affiliations: [1] Institute of Epidemiology and Social Medicine, University of Münster, Münster; [2] Department of Medical Biometry and Epidemiology, University Medical Center Hamburg-Eppendorf, Hamburg

Introduction

Clustered diagnostic test accuracy (DTA) data arise when one individual contributes multiple test results, e.g. across several lesions or at repeated time points. These observations are typically correlated, and ignoring these dependencies can bias estimates and underestimate variance. This systematic review aimed to identify, categorise, and evaluate statistical methods for analysing DTA in clustered data settings, focusing on approaches that address temporal and spatial clustering when estimating sensitivity/specificity and area under the receiver operating characteristic (ROC) curve (AUC).

Methods

Following a pre-registered protocol, we searched MEDLINE, Web of Science, and Embase, supplemented by citation tracking and reference-list screening (11/2025). Eligible studies were methodological articles, simulation studies, applied DTA studies, or systematic reviews. Studies unrelated to diagnostics, without clustering adjustment, or insufficient methodological description were excluded. Two reviewers independently screened, assessed eligibility, and extracted data. Included records were charted by design, dependence structure, assessment level, clinical task, target metric, and method family.

Results

Of 5,250 retrieved records, 3,580 were screened after deduplication, and 280 full-text articles were assessed for eligibility. Among applied studies with clearly reported clustered DTA structures, within-subject dependence was not adequately handled in a substantial subset. Initial analyses show that the majority of articles are methodological, alongside many applied studies, some hybrid articles, and a few review and guidance papers. The clearest pattern was defined by the type of clustered diagnostic scenario rather than by statistical approach alone. Spatial settings were more common than temporal and differed in level of analysis and clinical use of testing. Accuracy measures were more evenly distributed, with studies targeting AUC/ROC and sensitivity/specificity occurring at similar frequency, and many addressed both. Likewise, nonparametric, parametric frequentist, and semiparametric approaches were all substantially represented, with no single modelling framework clearly dominating.

Discussion

The identified literature is heterogeneous and suggests that clustered DTA is not one generic modelling problem, but a set of scenarios that differ by dependence structure, level of assessment, and clinical use. This review indicates whether clustering is appropriately accounted for and provides an empirical basis for systematically comparing methods used to analyse clustered DTA studies and for deriving recommendations.

DFG project: 539658720 protocol DOI: 10.37766/inplasy2025.8.0006.

[30-P6-03] Simultaneous comparison of multiple F1-scores in diagnostic tests

Presenter: Kanae Takahashi

Authors: Kanae Takahashi [1]

Affiliations: [1] Osaka Metropolitan University, Osaka

Introduction

Diagnostic tests are often evaluated in terms of binary outcomes, where performance is commonly assessed using sensitivity, specificity, and predictive values. The F1-score, defined as the harmonic mean of precision (positive predictive value) and recall (sensitivity), has been increasingly used as a summary measure. While statistical procedures exist for comparing two F1-scores, practical settings often involve multiple diagnostic tests evaluated simultaneously. This study aims to develop a statistical testing procedure for comparing more than two F1-scores.

Methods

We consider multiple diagnostic tests with dichotomous outcomes and represent their joint distribution using a contingency table with cell probabilities. Using the multivariate central limit theorem and the delta method, we derive the asymptotic distribution of the vector of estimated F1-scores. Based on this, we construct a Wald-type test statistic to assess the null hypothesis that all F1-scores are equal. The framework is illustrated for three tests and can be extended to more general settings.

Results

A simulation study with 100,000 repetitions per scenario was conducted to evaluate the performance of the proposed test. Under the null hypothesis, the empirical type I error rates approached the nominal level as the sample size increased, although slight inflation was observed for small samples. Under the alternative hypothesis, the empirical power increased with sample size, indicating the ability to detect differences among F1-scores.

Discussion

The proposed method provides a framework for simultaneously comparing multiple F1-scores in diagnostic tests. It performs well in large samples, although caution is warranted for small sample sizes due to potential type I error inflation. Future work will extend the proposed framework to a simultaneous comparison of macro- and micro-averaged F1-scores.

[30-P6-04] Statistical Methods for Assessing Diagnostic Agreement in AI Validation Studies: A Practical Review

Presenter: Maximilian Pilz

Authors: Maximilian Pilz [1]

Affiliations: [1] Nürnberg School of Health, Ohm University of Applied Sciences Nuremberg, Nuremberg

Introduction

With the rise of artificial intelligence (AI), an increasing number of AI-based diagnostic tools are being developed. Before clinical implementation, these tools must be validated against existing gold standards. This requires trials that quantify the agreement between AI predictions and reference measurements. However, designing such agreement studies poses methodological challenges that differ substantially from classical superiority trials. This talk aims to provide practical guidance for researchers and clinicians on how to design, analyze, and report diagnostic agreement studies for AI-based tools.

Methods

A review was conducted to identify established statistical methods for assessing diagnostic agreement. Methods were categorized according to the type of data (continuous, categorical, ordinal) and the number of raters or measurements involved. For each method, we summarized assumptions, appropriate use cases, interpretation of results, and available open-source software for sample size calculation and analysis.

Results

Several established methods applicable to AI validation studies were identified: Cohen's kappa and extensions (weighted, Fleiss'), Gwet's AC, Krippendorff's alpha, intraclass correlation coefficient (ICC), Lin's concordance correlation coefficient, and Bland-Altman analysis. Key distinctions, such as when to apply absolute agreement versus consistency metrics, are clarified. Furthermore, we provide step-by-step guidance on sample size calculation, including formulas, software recommendations (e.g., R packages), and common pitfalls.

Discussion

Agreement studies are essential for validating AI-based diagnostic tools, yet methodological guidance remains often fragmented standard curricula. By consolidating recommendations for study design, method selection, and sample size planning, this talk supports researchers in conducting rigorous validation studies. Ultimately, transparent and well-designed agreement studies are critical to ensure that AI tools meet the standards required for safe clinical deployment.

[30-P6-05] A Domain-Based Quality Assessment Framework for Fetal Cardiac Reference Value Studies: Development, Piloting, and Inter-Rater Reliability in the Age of AI-Assisted Prenatal Diagnostics

Presenter: Anita C. Windhorst

Authors: Anna M. Nist [1]; Anita C. Windhorst [2]

Affiliations: [1] Institute of Medical Informatics, Justus-Liebig-University Gießen, Giessen; [2] Institute of Medical Informatics, Justus-Liebig-University Gießen, Gießen

Introduction

Fetal cardiac z-score reference values underpin prenatal diagnostics and serve as normative benchmarks for AI-assisted echocardiographic screening. Their clinical validity depends on the quality of the underlying normative datasets: small, gestational-age-structured cross-sectional studies for which no fit-for-purpose quality assessment instrument exists. QUADAS-2 targets diagnostic accuracy, RoB 2 intervention trials, and the NHLBI cross-sectional tool exposure-outcome relationships none apply to normative reference curve construction. Within a PROSPERO-registered systematic review (CRD420251249222), we developed a purpose-built framework.

Methods

We adapted QUADAS-2 into a Quality Assessment Tool for Reference Value Studies (QA-RVS) with four domains:

Reference Population: sample size per Royston, inclusion/exclusion criteria, representativeness, STROBE demographic reporting Measurement Quality: ISUOG/ASE protocol standardisation and intra/inter-observer reliability Statistical Methods: LMS, GAMLSS, fractional polynomials, quantile regression, and polynomial models with AIC/BIC-based selection are adequate mean $\pm$ SD without gestational-age modelling is not Reporting Quality: TRIPOD criteria for z-score transparency and online calculator availability. Criteria were classified as critical or supporting with hierarchical risk derivation (Low/Moderate/High). Two raters assessed 10/59 studies independently Cohen's unweighted κ was calculated per criterion.

Results

Three criteria achieved excellent to perfect agreement: inclusion/exclusion criteria (κ=1.00), z-score reporting (κ=1.00), and statistical model adequacy (κ=0.82). Overall κ across 16 individual criteria was 0.52 (67.5% raw agreement), median 0.32 (range: -0.04-1.00) 3 of 7 critical criteria achieved $\kappa \geq 0.60$. Reliability was lower for measurement quality criteria, consistent with their higher information density. Study-level quality ratings agreed in 2 of 10 cases, driven by unresolved operationalisation of the Unclear category in domain-level derivation rules. Sources of disagreement have been identified and categorised.

Discussion

These are preliminary results from an ongoing review of 59 studies. Full inter-rater reliability data following a consensus meeting and complete quality assessments will be presented at the conference. The pilot demonstrates that agreement on conceptually central criteria is achievable from the outset. Lower reliability for measurement quality criteria reflects their higher information demands rather than conceptual ambiguity. The framework is transferable beyond fetal echocardiography to any domain where z-score reference values serve as clinical or computational benchmarks.

[30-P6-06] How was the trade? A review of the COVID-19 pandemic effects on global trade using a gravity model approach

Presenter: Anita C. Windhorst

Authors: Daniel Leufkens [1]; Jörn Pons-Kühnemann [2]; Anita C. Windhorst [3]

Affiliations: [1] Institute of Medical Informatics (IMI), Justus Liebig University Giessen, Giessen; [2] Institute of Medical Informatics (IMI), Justus Liebig University Giessen, Gießen; [3] Justus-Liebig-University Gießen, Gießen

Introduction

The COVID-19 pandemic that broke out in 2019 was undoubtedly an novum event in human history. About 7 million people died as a result of this global pandemic (World Health Organization (WHO), 2026). Health and economic systems around the world were strained to an extent never seen before (International Monetary Fund (IMF), 2020). In particular, international trade came under extreme pressure, as containing the pandemic was the top priority, at the same time health and economic systems could not do without global trade. In this regard, one of the main questions that arises is the impact of the COVID-19 pandemic on trade, to learn from this experience and be better prepared for future pandemics.

Methods

This article therefore models global trade using a gravity model commonly applied in trade theory (Tinbergen, 1962 Anderson and van Wincoop, 2003 Frimpong et. al., 2023) and uses this model (including Poisson pseudo-maximum likelihood) to examine the global trade effects of COVID-19 incidence and vaccination rates. For this purpose, quarterly trade data from 246 countries for the 24 quarters from 2019 to 2024 were collected (i.e., UNcomtrade, World Bank, IMF, CEPII) and supplemented with COVID-19 data provided by the WHO (2026).

Results

The results of the gravity panel regressions indicate that the rise in incidence and vaccination rates led to an increase in global trade. Over time, the pandemic initially caused a significant downturn in global trade before a vaccine became available however, this downturn came to a halt with the introduction of the COVID-19 vaccine, and trade subsequently rose significantly again.

Discussion

Thus, the COVID-19 pandemic does not appear to have led to a general downturn in global trade, although it is, of course, open to debate how global trade would have developed in the absence of the pandemic. Consequently, it is plausible to conclude that the introduction of a vaccine has helped mitigate, if not actually overcome, a recession in global trade.

[30-P6-07] Sex differences in diagnostic performance of non-invasive scores for steatotic liver disease (SLD)

Presenter: Martyna Flis

Authors: Martyna Flis [1]; Marc-Nicolas von Itter [1]; Tobias Nonnenmacher [2]; Tobias Norajitra [3]; Fabian Bamberg [1]; Christopher Schlett [1]; Susanne Rospleszcz [1]

Affiliations: [1] University Medical Center Freiburg, Department of Radiology, Freiburg; [2] University Hospital Heidelberg, Department of Diagnostic and Interventional Radiology, Heidelberg; [3] German Cancer Research Center (DKFZ), Division of Medical Image Computing (MIC), Heidelberg

Introduction

Steatotic liver disease (SLD) is characterized by excess fat accumulation in the liver and its diagnostic gold standard is medical imaging. This is not always feasible for screening, so several non-invasive tests (NITs) have been developed to diagnose SLD. However, given sex-specific differences in the underlying markers used in NITs and in SLD prevalence, their diagnostic performance may differ between men and women. In this study, we evaluate the performance of commonly used NITs and their cutoff values in men and women and assess whether sex-specific cutoff values may be warranted.

Methods

We used cross-sectional data from the baseline examination of the German National Cohort (NAKO), including N=12,633 women (mean age 49±12 years) and N=16,050 men (48±12 years). The following NITs were calculated, based on anthropometric and laboratory data: Fatty Liver index (FLI), hepatic steatosis index (HSI), and lipid accumulation product (LAP). SLD was defined as proton density fat fraction > 5.75% based on hepatic fat content quantified by multi-echo Dixon magnetic resonance imaging (MRI). Overall discriminatory ability of the indices, and performance according to sex-specific cutoffs were assessed using areas under the receiver operating characteristic curve (AUC).

Results

Using lower rule-out thresholds (FLI <30 and HSI <30) resulted in high sensitivity but low specificity in both men and women, whereas higher rule-in thresholds (FLI ≥60 and HSI ≥36) yielded a more balanced performance. Overall discrimination was good, with AUCs ranging from 0.74 to 0.84 in men and from 0.80 to 0.88 in women. FLI showed the highest AUC in both sexes. Sex-specific differences in optimal cutoffs were observed for FLI and lipid accumulation product (LAP) (FLI men: 62.6 FLI women: 42.6). Other indices showed more similar thresholds between sexes.

Discussion

Non-invasive liver-related scores showed good performance for diagnosing SLD, with slightly better discrimination in women. FLI demonstrated the highest diagnostic accuracy. Sex-specific differences in optimal cutoffs were observed for FLI and LAP, suggesting that sex-specific thresholds may improve the clinical utility of these tests. Overall, these findings suggest sex differences in the diagnostic performance and optimal cutoffs of some NITs for detecting SLD.

[30-P6-08] Assessment of insulin resistance based on a cut-off for SHBG in relation to the triglyceride-glucose index (TyG), TyG-BMI, and TyG-WC indices in women with PCOS

Presenter: Aleksander Owczarek

Authors: Aleksander Owczarek [1]; Marta Kochanowicz [2]; Tahan Ben [2]; Mariusz Wójtowicz [3]; Jerzy Chudek [4]; Magdalena Olszanecka-Glinianowicz [5]

Affiliations: [1] Medical University of Silesia in Katowice, Poland, Katowice; [2] Department of Gynecological Endocrinology, Faculty of Medical Sciences in Katowice, Medical University of Silesia in Katowice, Katowice; [3] Clinical Department of Gynecology and Obstetrics, Faculty of Medical Sciences in Zabrze, Medical University of Silesia in Katowice, Zabrze; [4] Department of Internal Medicine and Oncological Chemotherapy, Faculty of Medical Sciences in Katowice, Medical University of Silesia in Katowice, Katowice; [5] Health Promotion and Obesity Management Unit, Department of Pathophysiology, Faculty of Medical Sciences in Katowice, Medical University of Silesia in Katowice, Katowice

Introduction

Insulin resistance (IR) is one of the key factors of the pathogenesis of polycystic ovary syndrome (PCOS). IR can be diagnosed with comparable accuracy based on the HOMA-IR values > 2.1 and SHBG levels £ 41.5 nmol/L. This study aimed to assess the consistency of SHBG threshold with cut-off values that discriminate IR based on the triglyceride-glucose (TyG) index, as well as the TyG-BMI and TyG-WC indices.

Methods

Anthropometric measurements and serum levels of glucose, triglycerides, insulin, and SHBG were retrieved from the medical records of 264 Caucasian women diagnosed with PCOS. The weighted mean SHBG level for IR, based on TyG and related indices, was calculated using receiver-operating characteristics (ROC) analysis combined with the meta-analysis method (summary ROC).

Results

The weighted mean cut-off value for SHBG assessing IR was less than 43.1 nmol/L (95% CI: 37.0-49.2). The pooled mean prevalence of IR across all indices was 36.1% (95% CI: 33.5 - 38.7%). The pooled sensitivity and specificity of SHBG levels were 74.7% and 66.9%, respectively, with a diagnostic odds ratio of 5.96 (95% CI: 3.63 - 9.80). In bivariate analysis, there was greater heterogeneity

in sensitivity (I2 = 62.6%) than in specificity (I2 = 38.3%). Despite the heterogeneity in both dimensions, the generalized bivariate heterogeneity I2 was only 1%. The AUC for the summary ROC curve was 0.733. The subgroup of women with SHBG levels below the aforementioned cut-off was characterized by significantly higher body mass, BMI, and waist circumference, serum insulin and triglyceride levels, as well as more frequent HOMA-IR > 2.1 and hypertriglyceridemia.

Discussion

The study proved that SHBG levels can serve as a marker of insulin resistance in Caucasian women with PCOS. A value below 43 nmol/L, determined via meta-analysis statistical methods, with high sensitivity and specificity, enables the detection of insulin resistance and requires close monitoring of liver steatosis and carbohydrate metabolism.

[30-P6-09] Psychological Impact of Paediatric Type 1 Diabetes Screening: Preliminary Results from the Italian Cohort of the EDENT1FI Study

Presenter: MariaGiovanna Scarale

Authors: MariaGiovanna Scarale [1]; Francesca Gatti [2]; Bianca Maria Milazzo [3]; Olimpia Vincentini [4]; Flavia Pricci [4]; Gaia Perego [2, 5]; Francesca Milano [5]; Valentina Di Mattei [2]; Clelia Di Serio [1, 6]; Olga Boiko [7]; Parth Narendran [8]; Emanuele Bosi [9, 10]

Affiliations: [1] University Centre for Statistics in the Biomedical Sciences (CUSSB), Vita-Salute San Raffaele University, Milan; [2] School of Psychology, Vita-Salute San Raffaele University, Milan; [3] General Medicine, Diabetes and Endocrinology Department, IRCCS San Raffaele Scientific Institute, Milan; [4] Istituto Superiore di Sanità, Rome; [5] Clinical and Health Psychology Unit, IRCCS San Raffaele Scientific Institute, Milan; [6] Biomedical Faculty, Università della Svizzera Italiana, Lugano; [7] Department of Applied Health Research, College of Medicine and Health, University of Birmingham, Birmingham; [8] Institute of Immunology and Immunotherapy, College of Medicine and Health, University of Birmingham, Birmingham; [9] Vita-Salute San Raffaele University, Milan; [10] Head physician of the General Medicine, Diabetes and Endocrinology Department and Director of the Diabetes Research Institute at Ospedale San Raffaele, Milan

Introduction

Paediatric screening for type 1 diabetes (T1D) is increasingly being implemented across Europe, yet evidence on its psychosocial impact within specific national contexts remains limited. Within EDENT1FI project [1], this study presents preliminary findings from the Italian cohort, focusing on parental psychological wellbeing following screening.

Methods

Parents/caregivers of children undergoing autoantibody T1D screening in Italy (within the D1Ce study [2]) were assessed at baseline (T1, prior to screening) and follow-up (T2, upon receipt of screening results). In this preliminary sample, only parents/caregivers of children with negative screening results were available. Psychological wellbeing was measured using the Hospital Anxiety and Depression

Scale (HADS), while health-related quality of life was evaluated through selected EQ-5D-5L dimensions. Analyses were restricted to paired observations. Changes over time were assessed using paired Wilcoxon signed-rank tests, with significance set at $p<0.05$.

Results

A total of 49 participants were included for HADS and 46 for EQ-5D-5L. Significant reductions in psychological distress were observed between T1 and T2. Median HADS anxiety scores decreased by 3 points ($p<0.001$), and depression scores by 1 point ($p=0.01$), indicating improved parental emotional wellbeing over time. No statistically significant changes were found in EQ-5D-5L dimensions. Median differences for anxiety/depression and pain/discomfort were 0 ($p=0.5$ and $p=1$, respectively), suggesting stability in perceived health status.

Discussion

These preliminary findings, based on caregivers of children with negative screening results, indicate no adverse psychological impact of paediatric autoantibody T1D screening and suggest improvements in parental wellbeing over time, supporting its acceptability in the Italian context. Data collection is ongoing across European centres, including longitudinal follow-up of caregivers of screen-positive children across five timepoints.

[1] Boiko O. et al. (2026) Exploring the psychosocial implications of general population screening for paediatric type 1 diabetes in Europe: protocol for a mixed-methods study. Diabetic Medicine. In press. doi:10.1111/dme.70297

[2] Vincentini O, Pricci F et al. (2025) Study protocol of D1Ce Screen: A pilot project of the Italian national screening program for type 1 diabetes and coeliac disease in the paediatric population. PLoS One 20(8): e0328624. https://doi.org/10.1371/journal.pone.0328624

[30-P6-10] Updated statistical review of regulatory requirements for AI diagnosis

Presenter: Naoki Ishizuka, Taro Shibata

Authors: Naoki Ishizuka [1]; Taro Shibata [2]

Affiliations: [1] Kyoto University, Kyoto; [2] National Cancer Center, Tokyo

Introduction

Many research of AI application for healthcare have been reported. However, in order to use these techniques in practice it is necessary to get regulatory approval from the authority in each country [1]. We have reviewed the latest status so called SaMD: Software as Medical Device in regulation for AI diagnosis with Gastrointestinal endoscopy as an example.

Methods

We identified the SaMD with AI/ML which has been already approved both Japan and US.

Then we review the data package including the design.

Endpoints Sample size Control group, or any comparison with humans Prospective or Retrospective Intervention or Observational

Results

We reported in the ISCB 2025 and updated result will be reported in the conference.

So far, according to the package inserts in Japan, test data were collected retrospectively and there was no prospective study conducted in all of thirteen SaMD. Performance evaluation tests were conducted while primary endpoint was sensitivity and secondary endpoint was specificity and 95% confidence intervals were reported. There are two SaMD that were evaluated in the comparison with human experts or non-experts, however the package insert for them do not show how many clinicians were attended for this performance test. Other than the above two SaMD, there were no simultaneous control group in their performance evaluation tests.

According to the summaries of FDA's 510(k) Premarket Notification, test data includes both retrospective performance evaluation tests concerning sensitivity and specificity and prospectively randomized controlled trials which primary endpoints are detection rate in terms of superiority and positive predictive rate as non-inferiority.

Discussion

There is NO universally common regulation to get approval as well as what to do in post marketing. The regulation for AI diagnosis is different in each country and there is no harmonization.

[1] Walradt T, Glissen Brown JR, Alagappan M, Lerner HP, Berzin TM.J. Regulatory considerations for artificial intelligence technologies in GI endoscopy. Gastrointest Endosc. 202092(4):801-806.

[30-P6-11] Agenda of the novel Special Interest Group for Medical Devices - Finding similarity within the diversity of medical devices

Presenter: Wilmar Igl

Authors: Wilmar Igl [1, 2]; Emilie Gérard [3, 2]; Martin Fahy [4, 2]; Mader Michael [5, 2]

Affiliations: [1] VERANEX, Uppsala, Sweden; [2] SIG Medical Devices (PSI, EFSPI), Sheffield; [3] Brain&Mind, Paris; [4] JenaValve, New Yok; [5] Nobel Biocare, Zürich

Introduction

There are fundamental differences in the scientific, clinical, and regulatory pathways for the development of medical devices relative to drugs and biologics. Therefore, the Special Interest Group for Medical Devices (SIG MedDev) as part of the PSI and EFSPI organizations was formed in December 2025 to study the implications of these differences for the application of statistical methods.

Methods

Motivation

Scientific. Medical devices cover a broad range of mechanisms of action, specifically their mechanical and electronic functioning. However, these aspects usually do not play a relevant role for drugs/biologics.

Clinical. The performance of a medical device is usually the result of a combined effect of medical device, procedure, and user (e.g., patient/clinician). However, the effects of the procedure or the user (e.g., usability/learning) usually do not have the same importance for drugs/biologics.

Regulatory. The required evidence for regulatory decision-making for medical devices is based on a different regulatory framework (i.e., ISO standards) than the development of drugs and biologics (i.e., ICH guidelines) and depends on device class, development stage, and target audience.

Results

Topics of Interest

Scientific. Selected endpoints and their specific statistical properties, eg distribution, associations or causal roles, for various, relevant domains, esp. mechanical and electronic functioning, will be discussed and recommendations for best practices for data generation and data analysis will be provided.

Clinical. The publication of the estimand framework in the ICH E9 (R1) guideline in 2020 has had a major impact on the development of drugs and biologics. The adaption of the estimand framework for medical devices is expected to be of similar value, e.g., considering device-specific intercurrent events, ie, mechanical or electronic failure, and corresponding strategies.

Regulatory. Common and specific requirements depending on regional legislations and institutions, e.g., EMA approval (CE) in the EU vs. FDA approval (510K, de novo, PMA) in the USA, will be identified and recommendations for best practices, especially for international trials, will be given.

Discussion

In summary, after the foundation of the SIG Medical Devices, a forum is now available for interested members of the (statistical) community to discuss and learn about high-impact topics with special relevance for medical devices.

[30-P6-12] Evaluation of the European Health Technology Assessment landscape and suggestion of a common framework to assess AI-based digital medical devices

Presenter: Sandrine Boulet

Authors: Sandrine Boulet [1]; Erick Tavares [1]; Yannick Binois [1, 2]; Judith Abecassis [3]; Corinne Collignon [2]; Louisa Stüwe [4]; Vincent Thévenet [5, 6]; Moreno Ursino [1]; Gaël Varoquaux [3]; Rodolphe Thiebaut [5, 6, 7]; Sarah Zohar [1]

Affiliations: [1] Inserm, Université Paris Cité, Inria, UMR 1346, HeKA, Paris; [2] Haute Autorité de Santé (HAS), Saint-Denis; [3] Inria, Soda, Palaiseau; [4] Digital Health Delegation, French Ministry of Health, Paris; [5] Université de Bordeaux, Inserm, Bordeaux Population Health Research Center, U1219, Bordeaux; [6] Service d'Information médicale, CHU de Bordeaux, Bordeaux; [7] Inria, SISTM, Bordeaux

Introduction

The integration of artificial intelligence (AI) into medical devices (MDs) can improve diagnostics and prognostics, enable personalized treatments, and optimize clinical workflows. Health Technology Assessment (HTA) bodies play a central role in informing regulatory, policy and reimbursement decisions. However, their evaluation approaches do not fully capture the distinctive characteristics of AI-MDs. These gaps may hinder adoption, increase the burden on decision-makers, and potentially expose patients to risks. In this context, our objective is to examine the extent to which the European HTA landscape is equipped to assess AI-MDs and propose a common framework to evaluate them.

Methods

(i) Mapping current European pathways for AI-MDs and their associated evidence requirements (ii) comparing European HTA approaches and (iii) proposing a "European AI-HTA framework".

Results

The European landscape is overall active but heterogeneous in terms of regulatory pace, initiatives to reimburse and adopt of AI-MDs. To address the various challenges arising from AI characteristics of certain MDs, a few countries and organisations have developed specific approaches that bring a piece of solution, such as, (i) the French HAS descriptive grid for AI-MDs, (ii) the WHO's structured documentation "Model Facts", (iii) the UK's NICE evidence standards framework and the Belgian mHealth to adjust the evidence to risk, (iv) fast-track pathways from Germany and France to meet the need for agility of HTA frameworks, (v) the EU AI Act's Change Management Plan and methodological frameworks like REVAFT (Repeating Validation and Fine-Tuning) to manage the dynamic lifecycle and (vi) third-party audited evidence and certifications to ensure confidence in the entire AI-system employed by the MD.

Discussion

HTA and regulatory approaches alone are insufficient without practical framework to support their implementation, in particular for AI-MDs. To this end, a common European AI-HTA framework must be further developed and integrated by HTA agencies, enabling a modular and adaptive HTA approach, while balancing flexibility with effective risk control. Harmonization of these approaches is also essential to ensure convergence across the European landscape. Finally, the proposed framework could be used to support implementation of appropriate national or European legal frameworks such as the HTA regulation.

[30-P6-13] Comparative Effectiveness of Nature-Based Social Prescribing to Alleviate Loneliness in Spain: A Decision-Analytic Model with Probabilistic Sensitivity Analysis

Presenter: Veronika Papon

Authors: Veronika Papon [1]; Beate Jahn [1]; Gaby Sroczynski [1]; Jill S. Litt [2, 3, 4]; Montse Masó-Aguado [5, 6]; Marc Olivella Cirici [7]; Marjan Arvandi [1]; Jùlia Santamaria Navarro [1]; Igor Kuchin [1]; Sibylle Puntscher [1]; Uwe Siebert [1, 8, 9]; Ursula Rochau [1]

Affiliations: [1] Institute of Public Health, Medical Decision Making and Health Technology Assessment, Department of Public Health, Health Services Research and Health Technology Assessment, UMIT TIROL - University for Health Sciences and Technology, Hall in Tirol; [2] Institute for Global Health (ISGlobal), Barcelona Biomedical Research Park (PRBB) Doctor Aiguader, Barcelona; [3] CIBER Epidemiología y Salud Pública (CIBERESP), Madrid; [4] Universitat Pompeu Fabra (UPF), Barcelona; [5] Research Group On Methodology, Methods, Models and Outcomes of Health and Social Sciences (M3O), Faculty of Health Sciences and Welfare, Centre for Health and Social Care Research (CESS), University of Vic-Central University of Catalonia (UVic-UCC), Barcelona; [6] Institute for Research and Innovation in Life Sciences and Health in Central Catalonia (IRIS-CC), Vic; [7] General Sub-Directorate for Addictions, HIV, STIs and Viral Hepatitis. Public Health Secretariat. Government of Catalonia, Barcelona; [8] Center for Health Decision Science, Departments of Epidemiology and Health Policy & Management, Harvard Chan School of Public Health, Boston; [9] Institute for Technology Assessment and Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston

Introduction

Loneliness is associated with adverse physical and mental health outcomes[1]. Nature-based social prescribing (NBSP) is an intervention proposed in the RECETAS randomized controlled trial (RCT) to alleviate loneliness[2]. To estimate the effectiveness of NBSP beyond the trial time horizon, we developed a decision-analytic model to inform healthcare and public health decision-making by synthesizing evidence on health outcomes. Given substantial uncertainty in model parameters arising from limited data, heterogeneity, and measurement error, a probabilistic sensitivity analysis was performed to test robustness of our results.

Methods

A Markov state-transition model was constructed to compare NBSP to usual care (UC) in Spanish older adults (60+) experiencing loneliness. Life years (LYs), quality-adjusted LYs (QALYs) and loneliness-free years (LFYs) were estimated over a lifelong time horizon. Key model inputs, such as transition probabilities and utilities, were derived from the RECETAS RCT in Barcelona, Spain, supplemented by data from published literature. Parameter uncertainty was addressed using appropriate probability distributions (i.e., beta for utility parameters and transition probabilities, lognormal for relative risks, hazard and odds ratios). A probabilistic sensitivity analysis was conducted using Monte Carlo simulations with 10,000 iterations. 95% credible intervals (CrI) were estimated for model-generated outcomes.

Results

Across 10,000 Monte Carlo simulations, NBSP yielded 24.11 (CrI:23.13-25.70) LYs, 15.30 (CrI:14.27-16.64) QALYs, and 11.46 (CrI:5.88-17.83) LFYs, whereas UC resulted in 23.71 (CrI:23.01-24.69) LYs, 14.79 (CrI:14.15-15.55) QALYs, and 8.06 (CrI:6.02-9.97) LFYs. NBSP increased LYs by 0.40 (CrI:-0.20-1.34), QALYs by 0.51 (CrI:-0.25-1.51), and LFYs by 3.40 (CrI:-1.65-9.37) compared to UC. NBSP was identified as the preferred strategy in 90% of all simulations when evaluating QALYs, LYs, or LFYs.

Discussion

We performed a probabilistic sensitivity analysis within our decision-analytic model to provide a comprehensive assessment of uncertainty, supporting more informed decision-making. Our results showed that the evaluated health outcomes remained robust in almost all simulations, favoring NBSP over UC.

[1]World Health Organization, From loneliness to social connection - charting a path to healthier societies: report of the WHO Commission on Social Connection. 2025: Geneva

[2]Lavelle Sachs A et al. Connecting through nature: A systematic review of the effectiveness of nature-based social prescribing practices to combat loneliness. Landscape and Urban Planning, 2024. 248.

[30-P6-14] The impact of measurement error in evaluation measures on cost-effectiveness analysis

Presenter: Satomi Okamura

Authors: Satomi Okamura [1]

Affiliations: [1] Japan Institute for Health Security, Tokyo

Introduction

When collecting utility values based on health-related quality of life for cost-effectiveness analyses in Japan, the EQ-5D-5L is recommended as the first choice. The EQ-5D-5L consists of five dimensions (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression), each rated by patients on a five-point scale. Such self-reported tools may contain measurement errors arising from the measurement method or respondent subjectivity. However, the impact of measurement error in the EQ-5D-5L on cost-effectiveness results has not been sufficiently examined. This study aims to quantitatively evaluate the impact of such errors on costs and cost-effectiveness measures by conducting simulations assuming measurement error in the EQ-5D-5L, based on existing cost-effectiveness analyses.

Methods

Measurement error was assumed to occur in two EQ-5D-5L dimensions: pain/discomfort and anxiety/depression. Four scenarios were defined by combining the degree of misclassification (high/low) and misclassification pattern (uniform/non-uniform), and data were generated for each scenario using Monte Carlo simulation. Cost-effectiveness analyses were conducted using two existing frameworks. The first is an in-trial economic evaluation based on a randomized controlled trial (Cockayne et al., 2021) examining fall prevention through home hazard assessments and environmental modifications by occupational therapists in older adults QALYs were calculated from utility values using the trapezium rule over a 12-month follow-up. The second is a framework (Yang et al., 2019) applying EQ-5D-5L primary data from a cross-sectional survey of end-stage renal disease patients to a

Markov model comparing hemodialysis and peritoneal dialysis. For each scenario, measurement errors were added to observed EQ-5D-5L values, utility values and QALYs were re-estimated, and the change and rate of change in the ICER were evaluated.

Results

Simulations showed ICER changes across all frameworks. In particular, the ICER changed substantially in the trapezium-rule-based framework under non-uniform misclassification scenarios.

Discussion

While measurement error in self-reported outcomes is widely recognized, its impact can extend to cost-effectiveness results, requiring careful consideration in interpretation and decision-making.

[30-P6-15] Cost-Effectiveness of timely surgery and timely inpatient rehabilitation for the management of hip fracture

Presenter: Zheng Jing Jimmy Hu

Authors: Zheng Jing Jimmy Hu [1, 2]; Walter Wodchis [1, 3]; Lehana Thabane [2, 4, 5]; Gerhard Fusch [2]; Salhab El-Helou [2]; Lawrence Mbuagbaw [6]; Judith Versloot [1]

Affiliations: [1] Trillium Health Partners, Mississauga; [2] Departments of Medicine and Health Research Methods, Evidence, and Impact, McMaster University, Hamilton; [3] Institute for Health Policy, Management and Evaluation, University of Toronto, Toronto; [4] The Research Institute of St. Joe's Hamilton, Hamilton; [5] University of Johannesburg, Johannesburg; [6] McMaster University, Hamilton

Introduction

Real-world data is an important aspect of clinical and policy decision-making, and conducting health economic evaluations using real-world data requires utilizing advanced statistical techniques. Herein, we used multiple statistical approaches to evaluate the cost-effectiveness of providing timely surgery, inpatient rehabilitation, or a combination of both for the management of hip fracture patients in Ontario, Canada.

Methods

We developed a Markov model to estimate costs and quality-adjusted life-years over a 5-year horizon from the perspective of a public payer. Using variables available and derived from administrative data, we fitted two Weibull survival regression models to estimate the risk of mortality and a second fragility fracture, and the transition probabilities between health states. We fitted gamma regression models to estimate the cost accrued per health state. EQ-5D utility values were obtained from an international prospective observational study. We performed base-case and probabilistic sensitivity analysis using 5,000 simulations for willingness-to-pay thresholds from 0 to300,000 (CAD). We performed internal and external validation of our models.

Results

The incremental cost-effectiveness ratio for timely rehabilitation alone and for both timely surgery and rehabilitation compared to neither were 124,162 and83,714 per QALY, respectively, while timely surgery alone was an absolute advantage strategy. Five-year base-case costs ranged from 130,747 to156,848, and QALYs from 2.13 to 2.37. Timely surgery alone had the highest probability of being cost-effective up to a willingness-to-pay threshold of 128,000 per QALY, and the combination of timely surgery and inpatient rehabilitation had the highest probability of being cost-effective after a willingness-to-pay threshold of130,000 per QALY. The Brier scores of the survival models remained below 0.20, while the actual average costs per health state were higher than the predicted costs.

Discussion

The evaluation shows that ensuring timely access to surgery and reducing avoidable delays in surgery can improve patients' quality of life and reduce health system costs. Providing both timely surgery and inpatient rehabilitation yields greater QALYs but incurs a larger health systems cost. Statistically, the study demonstrates the use of regression methods for developing health economic models, but note that these models are limited by their parametric properties.

[30-P6-16] Evaluating First Aid Knowledge, Training Recency, and Emergency Preparedness Among German Teachers

Presenter: Johannes Schobel

Authors: Katharina Beyrich [1]; Peter Kuhn [2]; Johannes Schobel [2]

Affiliations: [1] Neu-Ulm University of Applied Science, Neu-Ulm; [2] DigiHealth Institute, Neu-Ulm University of Applied Sciences, Neu-Ulm

Introduction

Not every teacher is properly trained to be a first responder in emergency situations, however, everyone is required to provide first aid in case of emergencies. Education, the amount of first responders, and continuous training varies due to state-specific regulations, resulting in knowledge gaps between teachers, across schools and even states. The manuscript aims to discuss these differences and analyse the extent to which teachers have theoretical, school-specific first aid knowledge and which factors influence first aid competence.

Methods

A voluntary anonymous study with teachers was conducted. The online survey collected data on knowledge of possible emergency situations in a school context, subjective confidence in action-taking, and potential influencing factors such as professional experience, passed time since the last first aid course, or working as a first responder at schools.

Results

Science, technology, engineering and mathematics (STEM) teachers with more than five years' professional experience, as well as first responders, achieved significantly better results in the first aid questions.

Discussion

A positive correlation between subjective confidence in dealing with emergencies and achieved scores in the knowledge questions could be observed. No significant differences regarding federal state, type of school, professional experience, or schools with and without school first aid service were noted. Similarly, there were no significant differences in knowledge of resuscitation in children between primary school teachers and teachers at other types of schools. The results indicate that teachers in Germany have a moderate to good level of theoretical knowledge of first aid. Future research should examine the effectiveness of federal state-specific regulations as well as practical application skills and uncertainties regarding first aid in everyday school life.

P7 - AI and Biomedical Data

Poster area: P7

[30-P7-01] Towards synthetic cancer registry data

Presenter: Kevin Lehmann

Authors: Kevin Lehmann [1]; Klaas Dählmann [1]; Stefan Meisegeier [2]; Christian Lüpkes [1]; Andreas Hein [3]

Affiliations: [1] OFFIS - Institut für Informatik, Oldenburg; [2] Robert Koch-Institut, Zentrum für Krebsregisterdaten, Berlin; [3] Universität Oldenburg, Abteilung für Assistenzsysteme und Medizintechnik, Oldenburg

Introduction

German clinical cancer registry data is gathered to improve the quality of oncological care. Although compliant with the European GDPR, registries are reluctant to share this data because e.g. rare cancer cases could potentially be used to re-identify individual patients.

To enable data access for development of hypotheses and preliminary analyses, synthetic data generated from real registry data could be worthwhile. Afterwards, the minimally required data can be requested from the registries, streamlining the lengthy data acquisition process for themselves and the registry alike. The challenge lies in balancing all relevant relationships and distributions while avoiding any data privacy breaches.

Using a real dataset on breast cancer provided by the German Centre for Cancer Registry Data (ZfKD), this contribution therefore proposes and evaluates a preliminary method based on iterative proportional fitting (IPF) to synthesize cancer registry data into the ZfKD format.

Methods

Globally, our IPF-based method maintains the distribution of values of the original dataset. Locally, we ensure medical consistency between properties for each individual cancer case through custom rules. These rules are derived from both medical literature and analysis of the provided dataset. Optional properties, such as the existence of a metastasis, are currently generated in 50% of all cases. Presently, the method only generates patient and diagnosis data.

Results

We evaluated privacy-representativity trade-offs of the synthetic data using normalized uniqueness, nearest neighbor distance ratio (NNDR) and across all patient and diagnosis features. Compared to the original dataset, uniqueness ranged 0 to 0.011 (median 0, mean 0.002), NNDR 0 to 1 (median 0.001, mean 0.174) and JS-divergence 0 to 0.833 (median 0.447, mean 0.442).

Discussion

Because the synthetic data is based on the original distribution, the low values for uniqueness and NNDR are expected since similar values are used in both datasets. The JS-divergence's mean and median of roughly 0.44 likely result from deviations introduced through the flat 50% chance for optional properties. The results need to be evaluated in context, since low scores in a discrete and continuous set of values impact potential data privacy issues differently. Additionally, there is no known metric to reliably indicate a data privacy insuring dataset.

[30-P7-02] Evaluating the Validity of Virtual Patient Cohorts in Anesthesia

Presenter: Vanessa Barbet

Authors: Vanessa Barbet [1]; Marie-Noëlle Sallaberry [2]; Louise Baschet [1]; Jérôme Cartailler [3, 4]; Karim Guessous [5]; Yoann Elmaleh [5]; Bernard Delvaux [5]

Affiliations: [1] Horiana, 80 bis rue Paul Camelle, Bordeaux; [2] Horiana, SanFrancisco; [3] U942 MASCOT INSERM, Université Paris Cité, Paris; [4] Department of Anesthesiology and Critical Care Medicine, Hôpital Lariboisière paris, Paris; [5] Hôpital Privé Claude Galien, 20 Route de Boussy-Saint-Antoine, Quincy Sous Sénart

Introduction

Synthetic cohort generation models that are based on real-world data and preserve the relationships between clinical variables could help to address issues of access, confidentiality and limited dataset size.

This proof-of-concept study evaluates the feasibility and validity of generating synthetic patients from individual data in a real-world cohort previously used to assess the impact of chronic beta-blocker therapy on morphine consumption in the recovery room after hip or knee replacement.

Methods

The dataset of patients undergoing prosthetic lower limb surgery between 2011 and 2024 contained 8,339 patients.

Two models were used: (1) Bayesian Networks (BN) and (2) Tabular Variational Autoencoder (TVAE). We constructed a directed acyclic graph (DAG) to describe the relationships between demographic, clinical variables and perioperative pain management. We parameterized a semi-parametric hybrid BN, then sampled posterior probability distributions via Monte Carlo simulation. The TVAE was parameterized with 4 hidden layers.

The quality of the synthetic cohort was then examined according to the four standard dimensions: fidelity (similarity of distributions, consistency of relationships between variables), utility (ability to reproduce relevant statistical patterns), confidentiality (absence of real-patient duplication), and equity (correct subgroup representation).

Subsequently, the synthetic cohort's ability to reproduce the results of the initial analysis on the impact of beta-blockers on pain, estimated by postoperative morphine use, was assessed.

Results

TVAE showed limited data fidelity and produced unstable results. The synthetic data generated by BN showed a high degree of distributional similarity: approximately 90% of the variables had a total variation distance below 0.1. The results indicate no confidentiality issues and an excellent correlation between variables. Our model successfully reproduced almost all the results of the published study on associations between beta-blockers and morphine use.

Discussion

This 'proof-of-concept' study proposes a robust and reproducible methodology for generating and validating synthetic patients in anesthesia based on BN explicitly defined with clinical expertise. It raises several important methodological questions, notably the question of the minimum dataset required to train a reliable model and it highlights the potential value of augmented cohorts combining real and synthetic data to enhance the statistical power of analyses, particularly in under-represented subgroups.

[30-P7-03] Evaluation of Artificial Intelligence-Generated Synthetic Data for Clinical Research in Secondary Cardiovascular Prevention of Patients with Dyslipidemia.

Presenter: Alice Bonomi

Authors: Arianna Galotta [1]; Sebastiano Saccani [2]; Andrea Coser [3]; Pablo Werba [1]

Affiliations: [1] IRCCS Centro Cardiologico Monzino, Milano; [2] AINDO Spa, Triste; [3] Aindo Spa, Triste

Introduction

Access to high-quality clinical data is essential for advancing medical research and developing robust statistical and artificial intelligence models. However, strict privacy regulations and logistical constraints often limit timely access to real-world data (RWD). In this context, synthetic data - artificially generated datasets that preserve the statistical properties of original data without exposing sensitive patient information - represent a promising solution. This study explores the use of synthetic data in secondary cardiovascular prevention in patients with dyslipidemia, using two real-world datasets from Centro Cardiologico Monzino.

Methods

Given the relatively high dimensionality and limited sample size of the datasets, we developed an innovative hybrid data generation approach. This method integrates large language models (LLMs) to embed clinical domain knowledge into the generative process. Initially, a preliminary synthetic dataset was generated using GPT-4o-mini based on structured dataset descriptions. This dataset was then used to pretrain a neural generative model developed by Aindo, which was subsequently fine-tuned on real clinical data. This strategy ensured patient privacy while accommodating local computational constraints.

Results

Statistical analyses showed strong agreement between real and synthetic datasets, with only a small number of variables exhibiting significant differences. Predictive models trained on synthetic data successfully replicated key associations observed in the original datasets. In the statin-associated muscle symptoms (SAMS) use case, variables such as therapy modification and creatine phosphokinase levels consistently emerged as relevant predictors in both datasets. In the real dataset, therapy modification was associated with SAMS (OR = 0.10 95% CI: 0.04-0.12), while in the synthetic data the effect size was attenuated (OR = 0.28 95% CI: 0.15-0.56), with a 70% overlap in confidence intervals. For creatine phosphokinase, the odds ratio remained stable at 1.01 in both datasets, although the confidence intervals did not overlap (real: 1.01-1.01 synthetic: 1.00-1.01). Similarly, a generalized linear model (GLM) was used to assess the association between percentage LDL-C change and therapeutic intensity. In both datasets, pharmacological intensity was identified as the most significant predictor of LDL-C reduction (β = 16.4, SE = 1.5 for real data β = 6.1, SE = 1.4 for synthetic data).

Discussion

In conclusion, synthetic data proved to be a reliable proxy for real-world datasets in exploratory analyses and model development. Although some effect sizes were slightly attenuated, the main clinical relationships were preserved. Future research will focus on refining LLMs and implementing domain-specific training strategies to further improve the fidelity and applicability of synthetic data in healthcare.

[30-P7-04] Optimal transport synthesises infant cohort by predicting challenge responses from adult cohort

Presenter: Pratyasha Pal

Authors: Pratyasha Pal [1]; Nandini Gadhia [2]; Daniel O'Connor [3]; Michail Smyrnakis [2]; Barbara Camanzi [2]

Affiliations: [1] University of Oxford, oxford; [2] STFC Hartree Centre, Warrington; [3] University of Oxford, Oxford

Introduction

Controlled human infection model (CHIM) studies provide detailed longitudinal data across baseline, post-vaccination, and post-infection timepoints. However, these are typically not feasible in vulnerable populations such as children. We propose a computational framework that leverages these adult time-resolved data to infer corresponding infection responses in infants using optimal transport.

Methods

We represent immune responses as shifts between biological states in a shared latent space derived from transcriptomics data. Data from an adult cohort with baseline, post-vaccination, and post-infection measurements, and an infant cohort with baseline and post-vaccination data, were harmonised and embedded using PCA. Optimal transport was applied to align response patterns and infer how state transitions in adults relate to those in infants. A learned mapping was then used to project adult infection-induced changes into the infant setting, generating simulated infant post-infection profiles.

Results

The simulated infant responses remained consistent with the structure of real infant data while capturing directional changes observed in adult infection. Predicted responses showed strong internal consistency and were significantly different from random alignments. While gene-level agreement between adult and simulated infant responses was moderate, broader patterns were more consistent, indicating partial conservation of immune response dynamics. The simulated infant profiles suggested a response dominated by inflammatory activation, with comparatively weaker antiviral-like signalling.

Discussion

These findings demonstrate that optimal transport can potentially transfer perturbation dynamics across cohorts, enabling biologically plausible simulation of otherwise unobservable infant challenge responses. While the generated data should be interpreted as hypothesis-generating, this approach provides a scalable and ethically viable framework for augmenting paediatric datasets. Future work will focus on validation across additional cohorts and incorporating richer, nonlinear representations to improve biological fidelity.

[30-P7-05] Generative Adversarial Networks: A Reproducible Framework for Synthetic Control Arms in Clinical Trials

Presenter: Nicole Cizauskas

Authors: Nicole Cizauskas [1]; James Wason [2]; Svetlana Cherlin [3]

Affiliations: [1] Newcastle University, Newcastle; [2] Population Health Sciences Institute, Newcastle University, UK, newcastle upon tyne; [3] Population Health Sciences Institute, Newcastle University, newcastle upon tyne

Introduction

Synthetic control arms (SCAs) are control groups generated using external or historical data with characteristics comparable to those of the treatment group. SCAs are particularly valuable in randomised controlled trials (RCTs) where recruitment, cost, or ethical constraints limit the feasibility of conventional control arms, such as in rare disease or pediatric trials.

SCAs can be generated using a range of statistical approaches, from regression-based methods to advanced machine learning. Recent studies have explored the use of Generative Adversarial Networks (GANs) to generate synthetic patient populations for case studies. However, existing publications focus on model performance metrics and do not provide reproducible implementations of the GAN models used. As a result, researchers seeking to apply GAN approaches must develop models from scratch using general-purpose frameworks that are not tailored to clinical trial data.

Methods

We developed a reproducible framework for generating SCAs by customising a Conditional Tabular GAN (CTGAN) to accommodate clinical trial datasets. The model was adapted to recognise common clinical data structures and incorporates evaluation procedures relevant to trial settings, including k-anonymity assessment and propensity score matching comparisons.

To improve accessibility and reproducibility, the framework is publicly available through a GitHub repository and an accompanying website (https://n-cizauskas.github.io/SCOG-web/) that allows users to upload datasets and generate synthetic control arms with adjustable model parameters.

Results

The framework was evaluated across several case study datasets to assess model performance and usability. Results demonstrate that model performance is sensitive to dataset characteristics and hyperparameter selection. Unadjusted GANs frequently perform best in AUC, k-anonymity, standard mean difference, and other metrics.

Discussion

GAN-based approaches offer several advantages for generating SCAs, including the ability to model complex multivariate relationships, integrate multiple data sources, and generate flexible synthetic cohorts. However, model performance remains dependent on the quality and representativeness of the input data. Datasets with substantial missingness, bias, or poor alignment with the target population may reduce the reliability of generated controls.

By providing a reproducible framework and accessible implementation, this work aims to facilitate broader application of GAN developed synthetic control arms in clinical trials.

[30-P7-06] Generating virtual control patients in clinical trials by the use of mathematical tumor growth modeling

Presenter: Moritz Warneke

Authors: Moritz Warneke [1]; Carolin Herrmann [1]

Affiliations: [1] Heinrich-Heine-Universität Düsseldorf, Düsseldorf

Introduction

While the generation of virtual patients by using mathematical modeling of tumor growth has received increasing attention in recent years, model uncertainty and patient heterogeneity pose problems in data driven tumor modeling. Our goal is to address these issues by incorporating model selection methods into the modeling process to obtain virtual patients from multiple underlying growth models.

Methods

We fitted three multivariate differential equation models patient-wise to longitudinal non-small-cell lung cancer tumor volume measurements in 585 human patients (Ghaffari Laleh et al., 2022). Three model selection variants are considered: selecting the model with the lowest AIC for a given patient, assigning weights proportional to the model's likelihood and a hybrid approach, which assigns weights and removes models below a pre-specified threshold. For generating a virtual patient, the underlying model is selected first, based on prior selection proportions. Then the model parameters are randomly sampled from a vine-copula describing the multivariate parameter distribution of the data set. The virtual patients are compared to the observations by using a bootstrapped Cramér-von-Mises test and graphical methods. In a simulated clinical example, we compare tumor volumes after treatment between intervention and control group, where the latter consists of real and virtual patients.

Results

QQ-Plots and the Cramér-von-Mises-test suggest that the AIC-based selection approach yields the most accurate simulated data set. However, the underlying distribution of a model might be sparse if the model is rarely selected. The weighting approach utilizes all observations for every model but yields distorted simulated data. The hybrid approach is a compromise between the two approaches, utilizing more data points while producing less distortion. Regarding the clinical example, replacing parts of the control group with virtual patients yields endpoints with similar means but larger variance.

Discussion

The tumor models used in this work are a first step towards modeling virtual patients under model uncertainty in clinical trials using simple differential equation models. This work can be considered as a framework for future implementations of more complex tumor models for generating virtual patients.

References

Ghaffari Laleh, N., et al. (2022). Classical mathematical models for prediction of response to chemotherapy and immunotherapy. PLoS Computational Biology, 18(2).

[30-P7-07] Prompting Strategies for Coding in AI-Assisted Qualitative Data Analysis

Presenter: Mireya Diaz

Authors: Mireya Diaz [1]

Affiliations: [1] Case Western Reserve University, Cleveland

Introduction

Coding is a central step in many qualitative data analyses (QDA). Artificial intelligence (AI) has made inroads into the field with performance dependent on the prompting strategy used. This work examines the effectiveness of different prompting strategies on AI performance.

Methods

Coding and theme development were performed on abstracts from a 16-study meta-analysis examining the association between air pollution and preeclampsia incidence. It also examined it in a second related corpus containing fewer formal sources. Coding was performed with R, NVivo 14, and ChatGPT 4.1. Several prompting strategies in increasing degree of complexity (zero-shot, few-shots, and chain-of-thought -CoT) were evaluated regarding their accuracy against manual coding. Direct thematic coding with AI was also examined.

Results

Zero-shot prompting (with or without CoT) provided the best results for structured text. Requesting a full word-by-abstract matrix increases sensitivity. Supplementing CoT with general topic information and role playing made it the best strategy for more colloquial text. Directly inquiring thematic coding within AI also showed good performance. Some prompting strategies are more prone to hallucinations. The latter are associated with errors in the frequency counting process while coding.

Discussion

This work offers performance assessment of an AI tool for coding and theme development through a variety of prompting strategies combined with head-to-head comparison with popular QDA tools using readily available data. During user interaction, AI chatbots provide suggestions regarding potential next steps which may enhance results during coding. AI tools may require replication to ensure consistent findings and mitigate their probabilistic nature. Integrating results from at least two tools (AI and traditional) provides more comprehensive and accessible results for QDA.

The full paper pertaining to this abstract will be published in the GMDS journal "Medizinische Informatik, Biometrie und Epidemiologie (MIBE)."

[30-P7-08] Semantic Retrieval-Augmented Translation for Radiology Reports Using Small Language Models - Algorithm Development and Evaluation

Presenter: Daniel Reichenpfader

Authors: Fabio Dennstädt [1]

Affiliations: [1] University Hospital Bern, Bern University, Bern

Introduction

Accurate machine translation (MT) of clinical text is essential for multilingual patient care, cross-border data sharing, and research reproducibility, yet the use of commercial APIs for such tasks raises substantial privacy concerns. We present Semantic Retrieval-Augmented Translation (S-RAT), a lightweight, ontology-guided framework that integrates the Unified Medical Language System (UMLS) and SNOMED CT concept retrieval into large language model prompts to improve medical terminology fidelity without fine-tuning or external APIs.

Methods

Using 500 English radiology reports from MIMIC-CXR, we evaluated multiple S-RAT variants with open-source language models (gemma3, med42-v2, deepseek-r1, and gpt-oss) as well as a commercial frontier model (Gemini 3 Pro). We compared reference-free quality estimation metrics (COMET-src and GEMBA-DA-noref) with human annotation. Furthermore, we assessed the translation coverage of the German SNOMED CT edition for the specific use case of radiology report translation.

Results

S-RAT prompting did not consistently outperform standard prompting, and COMET-src correlated only weakly ($\rho \approx 0.24$, $p \approx 0.05$) with human adequacy judgments, while GEMBA-DA-noref showed no meaningful correlation. The German SNOMED CT edition covered only 32.6 % of extracted concepts.

Discussion

These findings highlight the current limitations of multilingual ontology resources and the application of general-domain MT metrics in clinical contexts. S-RAT demonstrates the feasibility of privacy-preserving, locally deployable, and ontology-augmented translation and provides a foundation for future work on domain-specific evaluation, ontology enrichment, and local deployment of language models in healthcare settings.

The full paper pertaining to this abstract will be published in the GMDS journal "Medizinische Informatik, Biometrie und Epidemiologie (MIBE)."

[30-P7-09] Interfaces for Agentic Clinical Decision Support Systems in a Precision Medicine Setting: are we there yet?

Presenter: Sara Nuhic

Authors: Sara Nuhic [1]; Stephanie Berger [2]; Stefan Sigle [1]

Affiliations: [1] MOLIT Institute gGmbH, Heilbronn; [2] SLK Kliniken Heilbronn, Heilbronn

Introduction

Clinical decision-making in precision medicine is complex due to the integration of multimodal clinical and molecular data [1] e.g. in molecular tumor board (MTB) workflows [2]. Current research on agentic Clinical Decision Support Systems (CDSS) emphasize model performance over workflow integration and data access. This work examines key data requirements and access mechanisms for agentic CDSS in MTB.

Methods

Literature on MTB related CDSS was reviewed identifying relevant data types and interfaces for agentic systems. Practical insights from the Virtual Tumorboard (VITU) [3] software were gathered by observing MTB case preparation and analyzing used data- and knowledge sources. Subsequently identified external knowledge sources were screened for agentic interfaces like Model Context Protocol (MCP) or Application Programming Interface (API).

Results

Case-specific reasoning and the identification of treatment options require both rich clinical context like diagnoses, disease stage, and prior treatments and structured molecular data, including genomic variants and their therapeutic actionability. External knowledge sources, such as biomedical databases, guidelines and trial registries, support evidence aggregation.

While some knowledge sources like CIViC and Zotero provide MCP interfaces, others like PubMed and ClinicialTrials.gov rely on APIs. Notably, MCP implementations are largely community driven.

Discussion

Agentic CDSS depend on both data quality and access. Integration of clinical, molecular, and external data remains challenging due to fragmented systems and limited machine-interpretable data, affecting reliable recommendations. Community-driven MCP implementations indicate demand for standardized access to biomedical data.

References

[1] D. Johnson, G. Del Fiol, K. Kawamoto, K.M. Romagnoli, N. Sanders, G. Isaacson, E. Jenkins, and M.S. Williams, Genetically guided precision medicine clinical decision support tools: a systematic review, J. Am. Med. Inform. Assoc. JAMIA. 31 (2024) 1183-1194. doi:10.1093/jamia/ocae033.

[2] A. Wahida, and R. Kurzrock, The Molecular Tumor Board Turns 10: The Age of Complexity, The Oncologist. 30 (2024) oyae271. doi:10.1093/oncolo/oyae271.

[3] C. Fegeler, D. Zsebedits, S. Bochum, D. Finkeisen, and U.M. Martens, Implementierung eines IT-gestützten molekularen Tumorboards in der Regelversorgung, Forum (Genova). 33 (2018) 322-328. doi:10.1007/s12312-018-0459-3.

Closing Session

Join us for the Closing Session as we look back on the conference, recognize outstanding contributions, and look ahead to next year's meetings.

Poster Awards

The session will begin with the Poster Awards, based on the results of the conference-wide Poster Voting. Throughout the conference, participants have helped select the winning posters by evaluating their scientific content, presentation and engagement. The awards will be presented by our LOC Member **Claudia Giuliani**.

Looking Ahead to 2027

The Closing Session will also offer a first look at what is planned for the next conferences. **Cécile Proust-Lima**, SPC Chair, and **Richard Post**, LOC Treasurer, will provide an outlook on ISCB 2027 in Rotterdam, The Netherlands. Conference presidents **Johanna Apfel-Starke** and **Oliver J. Bott** will then introduce GMDS 2027 in Hannover, Germany.

Recap and Farewell

To conclude the conference, SPC Chair **Nadine Binder** and LOC Chair **Harald Binder** will look back on the conference and officially close the meeting.

We hope you will join us for this final gathering as we bring the conference to a close and look forward to meeting again in 2027.

Mini Symposia

SmallData Symposium

Organizers

Harald Binder, Nadine Binder, Sophia Nolde, Marc Schumacher, Iván Acevedo Monterrosa, Angelika Rohde, Anna Köttgen, Rolf Backofen

Focus

Join our event for shaping the small data community to advance AI and modeling in settings with limited data - hosted by the Collaborative Research Center (CRC) 1597 "Small Data". Discover the forefront of small data innovation at the SmallData Symposium, October 1st, 2026. The event will feature experts within the field of small data, panel discussions, and poster sessions to foster interdisciplinary exchange and create a shared small data language.

Agenda

09:00-09:10 Harald Binder: Opening Remarks.

Session 1: Similarity.

09:10-09:30 Nadine Binder and Jan Hasenauer: AI-assisted continuous-time modeling of metastatic breast cancer under sparse and irregular observation.

09:30-09:50 Maria Kalweit and Evelyn Ullrich: BLINK: Modeling NK cell cytotoxicity as a latent dynamical process from time-resolved fluorescence microscopy.

09:50-10:30 Panel Discussion: The role of similarity in small data. Panelists: Nadine Binder, Jan Hasenauer, Maria Kalweit, Evelyn Ullrich.

10:30-11:00 Break.

Session 2: Transfer.

11:00-11:25 Frank Hutter and Pascal Schlosser: Advancing Tabular Foundation Models for longitudinal data integration.

11:25-11:50 Anna Köttgen and Johannes Hertel: From population studies to whole body modeling and back.

11:50-12:25 Early career researcher flash presentations on collaborative work. Presenters: Nils Kober, Adrian Fritz, Hanning Yang, Jelena Bratulic, Sebastian Walter, Carola Heinzel, Fabian Kabus.

12:25-13:30 Lunch and poster session.

Session 3: Uncertainty.

13:30-13:50 Heinz Wiendl and Susanne Weber: Clinical categories to statistical trajectories: An interdisciplinary approach to modeling disease progression in multiple sclerosis.

13:50-14:10 Martin Wolkewitz and Maria Elena Maccari: Challenges of target trial emulation in rare pediatric diseases: from limited data to clinical management.

14:10-14:50 Panel Discussion: The role of uncertainty in small data. Panelists: Heinz Wiendl, Martin Wolkewitz, Maria Elena Maccari, Peter Pfaffelhuber.

14:50-15:00 Harald Binder: Outlook on the small data landscape.

SmallData Symposium Session 1: Similarity

Thu, 01.10.26, 09:00–10:30, Room: Rolf Böhme Saal

Chair: Max Behrens

Session 1: Similarity.

09:10-09:30 Nadine Binder and Jan Hasenauer: AI-assisted continuous-time modeling of metastatic breast cancer under sparse and irregular observation.

09:30-09:50 Maria Kalweit and Evelyn Ullrich: BLINK: Modeling NK cell cytotoxicity as a latent dynamical process from time-resolved fluorescence microscopy.

09:50-10:30 Panel Discussion: The role of similarity in small data. Panelists: Nadine Binder, Jan Hasenauer, Maria Kalweit, Evelyn Ullrich.

AI-assisted continuous-time modeling of metastatic breast cancer under sparse and irregular observation

09:10-09:30

Presenter/Author: Nadine Binder, Jan Hasenauer

Affiliations: Medical Center - University of Freiburg, Freiburg im Breisgau; Bonn Center for Mathematical Life Sciences, University of Bonn, Bonn

Introduction

Biomedical decision-making often relies on small, heterogeneous, and irregularly sampled patient cohorts in which clinically relevant events are only partially observed. This is particularly true for metastatic breast cancer, where organ-specific progression unfolds over time, observation intervals vary across patients, and both biological similarity and uncertainty must be handled explicitly. We present an AI-assisted modelling framework for this setting that combines continuous-time multistate modelling with representation learning for complex treatment information.

Methods

Using longitudinal trajectories from a large multicentre real-world registry, we model metastatic progression as a continuous-time Markov chain on combinations of metastatic sites. This formulation accommodates irregular follow-up, supports patient-specific covariates, and quantifies how existing metastases alter the risk of subsequent spread to other organs. To represent high-dimensional and heterogeneous treatment histories in a data-efficient way, we derive low-dimensional similarity-based covariates from biomedical language model embeddings. Uncertainty is propagated through likelihood-based inference and Hamiltonian Monte Carlo, yielding credible intervals for subtype-specific organ tropism, covariate effects, and inter-organ interactions.

Results

The analysis reveals structured, non-independent progression patterns despite sparse observations, including subtype-specific dependencies between visceral metastases and later cerebral spread.

Discussion

More broadly, the work highlights how similarity-aware AI components and uncertainty-aware stochastic models can be combined for small-data biomedical analysis: language models help encode clinically meaningful similarity in complex covariates, while mechanistic probabilistic modelling preserves interpretability and quantifies uncertainty in downstream risk estimates. This provides a principled framework for risk stratification in settings where data are limited, observations are intermittent, and reliable uncertainty assessment is essential.

BLINK: Modeling NK Cell Cytotoxicity as a Latent Dynamical Process from Time-Resolved Fluorescence Microscopy

09:30-09:50

Presenter: Evelyn Ullrich, Maria Kalweit

Authors: Iman Nematollahi [1]; Jose Francisco Villena-Ossa [2]; Alina Moter [3]; Kiana Farhadyar [1, 4]; Gabriel Kalweit [5, 1]; Abhinav Valada [1]; Toni Cathomen [2]; Evelyn Ullrich [3]; Maria Kalweit [1, 5, 4]

Affiliations: [1] Department of Computer Science, University of Freiburg, Freiburg im Breisgau; [2] Institute for Transfusion Medicine and Gene Therapy, University Medical Center Freiburg, Freiburg im Breisgau; [3] Goethe University, Department of Pediatrics, Experimental Immunology and Cell Therapy, Frankfurt am Main; [4] IMBIT//BrainLinks-BrainTools, University of Freiburg, Freiburg im Breisgau; [5] Collaborative Research Institute Intelligent Oncology (CRIION), Freiburg im Breisgau

Introduction

Natural killer (NK) cell cytotoxicity arises from dynamic interactions with tumor cells and is commonly assessed using time-resolved fluorescence microscopy. While apoptotic events can be annotated at individual frames, cytotoxic outcome emerges over time and cannot be reliably inferred from frame-wise classification alone. This motivates modeling cytotoxicity as a temporal process driven by latent interaction dynamics.

Methods

We propose BLINK, a trajectory-based latent dynamical model for estimating cumulative NK-induced tumor cell death. The model is formulated as a recurrent state-space model that learns latent interaction states from multi-channel microscopy sequences under partial observability. Cytotoxic outcome is modeled via a monotonic increment formulation: a neural prediction head estimates

non-negative apoptosis increments that accumulate over time, ensuring biologically consistent cumulative predictions. The model is trained end-to-end by combining latent reconstruction, temporal consistency, and supervised outcome estimation.

Results

We evaluate BLINK on long-term NK-tumor co-culture recordings (~10 hours, ~250k frames). Compared to frame-based and recurrent baselines, BLINK improves trajectory-level cytotoxic outcome prediction (MAE 0.60 vs. 0.74 for the best baseline correlation 0.77 vs. 0.57) and substantially enhances forecasting accuracy (F-MAE30 0.05 vs. 0.22). Deterministic models without monotonic constraints collapse under sparse event supervision, whereas the proposed formulation remains stable. In addition, the learned latent space organizes NK trajectories into interpretable behavioral modes (e.g., cytotoxic, motile, quiescent) and reveals coherent temporal transitions consistent with known interaction phases.

Discussion

Modeling cytotoxicity as inference over latent interaction dynamics provides a principled alternative to frame-wise classification in time-resolved microscopy. The monotonic increment formulation enforces biologically meaningful constraints and improves robustness in sparse outcome settings. Beyond predictive performance, BLINK enables interpretable analysis of single-cell interaction behavior and supports forecasting of future outcomes. This framework offers a step toward quantitative, data-driven modeling of immune cell function and may facilitate scalable evaluation of cellular immunotherapies.

SmallData Symposium Session 2: Transfer

Thu, 01.10.26, 11:00–12:30, Room: Rolf Böhme Saal

Chair: Julia Hindel

Session 2: Transfer.

11:00-11:25 Frank Hutter and Pascal Schlosser: Advancing Tabular Foundation Models for longitudinal data integration.

11:25-11:50 Anna Köttgen and Johannes Hertel: From population studies to whole body modeling and back.

11:50-12:25 Early career researcher flash presentations on collaborative work. Presenters: Nils Kober, Adrian Fritz, Hanning Yang, Jelena Bratulic, Sebastian Walter, Carola Heinzel, Fabian Kabus.

Advancing Tabular Foundation Models for Longitudinal Data Integration

11:00-11:25

Presenter/Author: Pascal Schlosser, Frank Hutter

Affiliations: Institute for Epidemiology and Prevention, Medical Center - University of Freiburg, Freiburg., Freiburg; Johns Hopkins University Bloomberg School of Public Health, Baltimore; Institut für Informatik, Albert-Ludwigs-Universität Freiburg, Freiburg im Breisgau

Introduction

Tabular foundation models (TFMs) have advanced rapidly for classification and regression, yet their application to time-to-event survival analysis remains largely unexplored. A core obstacle is censoring: survival datasets routinely contain observations whose event times are only partially observed, and standard TFMs lack any native mechanism to incorporate this, producing biased estimates that limit their practical utility. We address this gap with \texttt{SurvPFN}, a prior-data fitted network (PFN) designed specifically for survival prediction.

Methods

Our approach casts survival as distributional regression under censoring and trains the model in-context on millions of synthetic survival datasets. Three ingredients make this possible: (i) a synthetic data-generating process based on Weibull-distributed event times paired with a non-informative censoring mechanism, (ii) explicit conditioning on a censored event indicator, and (iii) a training objective based on a censored negative log-likelihood. The resulting model performs survival prediction in a single forward pass without per-dataset training.

Results

We evaluate \texttt{SurvPFN} on SurvSet, a benchmark suite of real-world survival datasets spanning diverse domains and sample sizes. Across these tasks, \texttt{SurvPFN} matches or exceeds both classical and deep-learning survival baselines, despite requiring no per-dataset fitting, no bespoke survival architecture, and no hand-crafted feature engineering.

Discussion

These findings indicate that framing survival as continuous-time distributional regression with an appropriate censored loss is sufficient to bring the strengths of PFNs to time-to-event prediction. By eliminating the need for task-specific tuning, \texttt{SurvPFN} opens a path toward foundation-model-style survival analysis and broadens the scope of in-context learning to a clinically and scientifically important class of problems.

From population studies to whole body modeling and back

11:25-11:50

Presenter/Author: Johannes Hertel, Anna Köttgen

Affiliations: University Medicine Greifswald, GREIFSWALD; Institute of Epidemiology and Prevention, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg im Breisgau

Introduction

Population-scale integration of high-throughput molecular data with genome-wide genetic data enable hypothesis-generating screens to elucidate previously unrecognized gene functions. However, interpretation, contextualisation, and validation can be challenging, in particular in the rare variant setting where low variant carrier numbers complicate robust statistical inference. Here, we demonstrate how complementary knowledge-based modelling can be integrated with genome-wide association screens to generate a research loop that enables discovery and knowledge-based contextualisation of gene functions in human metabolism at the same time.

Methods

We utilised data from 4737 participants of the German Chronic Kidney Cohort with available whole exome sequencing as well as urine and plasma metabolomics data. For identifying gene-metabolite associations, we performed rare variant aggregation testing. Identified associations were then mapped onto sex-specific, organ-resolved whole-body models (WBMs) of metabolism and in silico knockout experiments were performed, solving linear and quadratic programming problems in the steady state solution spaces of the WBMs.

Results

We identified 235 significant gene-metabolite associations, many of which were not reported before. Orthogonal approaches, including experimental validations, confirmed results and added biological context. Moreover, we showed that heterozygous rare variant carrier status informed about consequences of inborn errors of metabolism (IEM), which result from homozygosity and were unobserved in the population because of their extremely low frequencies. Of 67 genes successfully mapped onto WBMs, in silico knockout predicted 78% of the observed metabolite associations correctly. An example for convergence of results from population analysis and modeling is 8-methoxykynurenine as a potentially novel biomarker of kynureninase deficiency, which was then validated in a clinical case of the resulting IEM. Microbiome-personalised WBMs showed promise in predicting empirical effect sizes of rare variants on metabolite levels as observed in the population analysis.

Discussion

Rare variant screens provide a powerful tool to investigate the genetic underpinnings of human metabolism with WBMs showing promise in providing biological context and in silico validation. Increasing the scope and depth of the introduced research loop between omics screens and metabolic modelling via the integration of proteome data on the population level and single cell omics data into the WBMs may enable a more fine-graded mapping of gene functions.

[01-P1-19] Convergence rate analysis of the nonparametric maximum likelihood estimator with external data

11:50-11:55

Presenter: Nils Kober

Authors: Nils Kober [1]; Max Behrens [2]; Harald Binder [2]; Angelika Rohde [1]

Affiliations: [1] Department of Mathematical Stochastics, University of Freiburg, Freiburg; [2] Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg

Introduction

Clinical prediction models commonly have to learn from small, heterogeneous biomedical datasets, while potentially useful external cohorts are collected following different sampling schemes or measurement conventions. In such settings, additional external data are valuable to the estimation on the target dataset only if their underlying distribution is similar in a suitable sense. We develop mathematical theory for the estimation of the cumulative distribution function of a random variable from current-status observations. To improve the accuracy for small target sample sizes in this classical model, we include observations from a larger external dataset with a similar but in general non-identical distribution.

Methods

We propose a weighted nonparametric maximum likelihood estimator that uses a convex combination of the two datasets in its criterion function. For desirable rates of convergence, the weights need to be chosen based on the Hellinger distance between the true underlying distributions of the observed data.

Results

We derive its rate of convergence in Hellinger distance, which exceeds the n^frac13-rate of the classical NPMLE for monotone functions if the external sample size grows faster than the target sample size and the external distribution converges suitably to the target distribution. Our main contribution is the derivation of maximal inequalities to bound the expected supremum of the weighted empirical process mathbbGnwn := wn 2-1 sumi=1n wn,i (deltaXi - mathbbPi) over a function class, where (Xi sim mathbbPi)i=1,dots, n are independent but not necessarily identically distributed and wn = (wn,1, dots, wn,n)^top are nonnegative, normalized weights.

Discussion

Currently, we only consider deterministic weights in the criterion function of the NPMLE. For real-world application of the method, they would need to be chosen adaptively based on observed data, which is an open research question so far. Our methodology is closely related to the work of Behrens et al., who developed a weighted parametric maximum likelihood estimator to fit local regression model for different patient subgroups. References [1] Kober, N. University of Freiburg. Master's thesis: Convergence rate analysis of the nonparametric maximum likelihood estimator with external

data. Not published. Jan 2026. [2] Behrens M, Stolz D, Papakonstantinou E, Nolde JM, Bellerino G, Rohde A, Hess M, Binder H. Contrasting Global and Patient-Specific Regression Models via a Neural Network Representation, Biometrical J. Mar 2026.

[01-P1-28] Benchmarking single-cell large language models

11:55-12:00

Presenter: Adrian Fritz

Authors: Adrian Fritz [1]; Niklas Brunn [2]; Jonatan Menger [3]; Anup Kumar [1]; Sonia Maria Krißmer [4]; Pavankumar Videm [1]; Ryan Köksal [1]; Maren Hackenberg [5]; Rolf Backofen [1]

Affiliations: [1] Department of Computer Science, Faculty of Engineering, University of Freiburg, Freiburg; [2] Institute of Medical Biometry and Statistics (IMBI), Faculty of Medicine and Medical Center, University of Freiburg, Germany, Freiburg; [3] Institute of Medical Biometry and Statistics (IMBI), Faculty of Medicine and Medical Center, University of Freiburg, Freiburg; [4] Medical Center Freiburg, Freiburg; [5] Institute of Medical Biometry and Statistics, University of Freiburg, Freiburg

Introduction

Foundation models are rapidly expanding in single-cell omics, but it remains unclear how language-based approaches compare with transcriptome-native models for biological interpretation. Recent work spans prompted general-purpose large language models, models that treat transcriptomic data as language, multimodal models combining text and quantitative information and expression-focused single-cell foundation models, yet these approaches are often evaluated in separate settings and under different assumptions.

Methods

We investigate this varying landscape through a comparative benchmarking framework with a particular focus on language-based and language-integrated foundation models for single-cell analysis. We center the comparison on cell type annotation as a practical and widely used benchmark task, while recognizing that cellular identity is often continuous and context-dependent rather than fully captured by discrete labels. Within a modular evaluation workflow that standardizes dataset specification, model wrapping, ontology-aware label harmonization and metric aggregation, we compare prompted generalist large language models, transcriptome-to-language models and multimodal text-cell models against expression-native single-cell foundation models and conventional annotation baselines. The benchmark is completely reproducible and publicly available as a nextflow workflow including both the model implementations and data sets. The modular setup facilitates easy extension to further data sets and models, not only for benchmarking purposes but even for usage on new data sets or for ensemble predictions.

Results

Across benchmark tasks, language-based and language-integrated models achieved reasonable annotation performance but did not consistently outperform transcriptome-native foundation models or established annotation approaches. In addition, prompted language models occasionally produced malformed, non-biological or contextually inconsistent labels, highlighting current robustness limitations in open-ended annotation settings.

Discussion

These findings suggest that language-oriented single-cell models are promising for flexible and multimodal biological interpretation, but currently complement rather than replace established transcriptome-centric methods. The presented benchmarking framework provides a reproducible basis for future evaluation of robustness, interpretability and open-ended language-driven analysis tasks in single-cell omics.

[01-P2-16] Agentic ODE Discovery and Parameter Distribution Inference from Summary Statistics for Rare Diseases

12:00-12:05

Presenter: Hanning Yang

Authors: Hanning Yang [1]; Meropi Karakioulaki [2]; Lennart Purucker [3, 4]; Tim Litwin [1]; Cristina Has [2]; Moritz Hess [1]

Affiliations: [1] Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center - University of Freiburg, Freiburg; [2] Department of Dermatology, Medical Faculty and Medical Center - University of Freiburg, Freiburg; [3] Prior Labs, Freiburg im Breisgau; [4] University of Freiburg, Freiburg, Germany, Freiburg

Introduction

Mechanistic modeling via ordinary differential equations (ODEs) offers interpretable descriptions of complex biological dynamics and is particularly valuable in rare disease research. However, rare diseases present a compounding set of challenges: model structure and parameters are typically unknown, individual-level data is scarce and noisy, and privacy constraints limit data sharing. These conditions demand methods that explicitly handle both structural uncertainty over which ODE governs the system, and parametric uncertainty over heterogeneous patient populations.

Methods

We present AgentODE, an end-to-end agentic framework that jointly discovers ODE structures and infers parameter distributions solely from population-level summary statistics. A large language model (LLM) proposes candidate ODE structures through an outer discovery loop, while a parameter inference agent iteratively refines parameter distributions via an inner diagnosis-update loop. At

each iteration, current parameter distributions are used to simulate synthetic trajectories. The agent then updates parameter distributions based on discrepancies between summary statistics from synthetic and empirical trajectories. An experience buffer with dynamic memory guides subsequent structure proposals, enabling structured exploration under uncertainty.

Results

We evaluate AgentODE on benchmark problems and two clinical datasets. One includes only 231 observations across 46 patients from a recessive dystrophic epidermolysis bullosa (RDEB) cohort . AgentODE achieves performance comparable to baselines with full individual-level data access, while operating exclusively on summary statistics. Critically, on RDEB, AgentODE recovers a mechanistically consistent ODE structure aligned with an independently developed expert model, whereas baselines, despite lower prediction error, recover implausible or degenerate structures. This reveals that in sparse, noisy settings, predictive performance alone is an insufficient criterion for mechanistic model quality.

Discussion

AgentODE demonstrates that population-level summary statistics, rather than being merely a privacy-preserving compromise, provide a more robust signal for uncertainty-aware structure discovery by condensing population patterns while mitigating individual-level noise. This opens a principled pathway for mechanistic modeling in rare diseases where data scarcity, heterogeneity, and privacy constraints have traditionally precluded such analyses.

[01-P2-07] Generating realistic bacterial microscopy from mechanistic simulators with limited real-world data

12:05-12:10

Presenter: Jelena Bratulić

Authors: Jelena Bratulić [1]; Moritz Steinmaier [1]; Jonas Pleyer [2]; Simon Ging [3]; Christian Fleck [2]; Thomas Brox [1]

Affiliations: [1] Computer Vision Group, University of Freiburg, Freiburg im Breisgau; [2] Freiburg Center for Data Analysis, Modeling and AI, University of Freiburg, Freiburg im Breisgau; [3] Computer Vision Group, University of Freiburg, Freiburg

Introduction

Microscopy-based downstream tasks such as cell segmentation and tracking rely on precisely annotated data, which is costly to obtain: experts must delineate cell boundaries by hand, and labeling lineages across cell divisions in video data is both laborious and inherently ambiguous. Agent-based simulators built on mechanistic cell and bacteria models, such as cr_mech_coli, bypass this annotation bottleneck by producing image-annotation pairs with pixel-perfect ground-truth masks, including cell positions, orientations, growth phases, and accurate lineage trees, under physically

plausible growth dynamics. However, naïve renders of simulator output appear overly synthetic for deep learning training, limiting generalization to real microscopy and constraining the practical use for large-scale synthetic data generation.

Methods

We build on the cr_mech_coli simulator for rod-shaped bacteria and present two approaches for closing the sim-to-real gap: a physics-inspired pipeline, crm_gen, and a learned style-transfer model, diff_gen. crm_gen renders synthetic micrographs across three modalities, bright-field, phase-contrast, and fluorescence, by composing independent modules that approximate sample absorption, optical point-spread-function blur, sensor noise, uneven illumination, and background structure. Custom modality styles are supported by automatically calibrating the modality-relevant module parameters via derivative-free optimization on a small reference set. diff_gen leverages a pretrained diffusion model fine-tuned with LoRA on a small reference set and a modality vector, conditioned either on the segmentation mask (style transfer only) or directly on the simulator's mechanistic state, replacing the rendering step.

Results

We evaluate both pipelines across all three modalities using image-similarity scores, distribution-shift measures, and qualitative comparison, and we analyze data robustness by varying the number of reference samples used for calibration and style transfer. Initial results indicate effective style transfer for both pipelines, even from small reference subsets.

Discussion

Classical calibration and learned diffusion-based style transfer close the sim-to-real gap through complementary mechanisms: crm_gen is fast and interpretable but requires per-style fitting, while diff_gen enables multi-style generalization at a higher resource cost. Although the gap is not closed in absolute terms, both pipelines demonstrate the value of agent-based simulators for generating large-scale synthetic data with accurate annotations, mitigating the need for expert labeling and supporting downstream training in small-data scenarios with limited real-world annotated data.

[01-P2-09] Using Knowledge Graphs to harvest datasets for efficient CLIP model training

Presenter: Sebastian Walter

Authors: Simon Ging [1]; Sebastian Walter [2]; Jelena Bratulić [3]; Johannes Dienert [3]; Hannah Bast [3]; Thomas Brox [4]

Affiliations: [1] University of Freiburg, Freiburg; [2] University of Freiburg, Freiburg im Breisgau; [3] Computer vision group, University of Freiburg, Freiburg im Breisgau; [4] Computer Vision Group, University of Freiburg, Freiburg im Breisgau

Introduction

Training high-quality CLIP models typically requires enormous datasets, which limits the development of domain-specific models – especially in areas that even the largest CLIP models do not cover well – and drives up training costs. This poses challenges for scientific research that needs fine-grained control over the training procedure of CLIP models.

Methods

In this work, we show that by employing smart web search strategies enhanced with knowledge graphs, a robust CLIP model can be trained from scratch with considerably less data.

Results

Specifically, we demonstrate that an expert foundation model for living organisms can be built using just 10M images.

Discussion

Moreover, we introduce EntityNet, a dataset comprising 33M images paired with 46M text descriptions, which enables the training of a generic CLIP model in significantly reduced time.

[01-P3-05] Improving Biogeographical Ancestry Prediction Through Machine Learning-Based Classification and Marker Selection

12:15-12:20

Presenter: Carola S. Heinzel

Authors: Carola S. Heinzel [1]; Lennart Purucker [2]; Julika Schmitz [3]; Jana Naue [4]; Theresa Maurer [1]; Zoe Lange [5]; Frank Hutter [6]; Sabine Lutz-Bonengel [3]; Holger Dette [5]; Peter Pfaffelhuber [1]

Affiliations: [1] Department of Mathematical Stochastics, University of Freiburg, Freiburg im Breisgau; [2] Machine Learning Lab, University of Freiburg, Freiburg im Breisgau; [3] Institute of Legal Medicine, Medical Center, University of Freiburg, Freiburg im Breisgau; [4] Institute of Forensic Medicine, Medical Center - University of Freiburg, Freiburg im Breisgau; [5] Ruhr-Universität Bochum, Bochum; [6] Institut für Informatik, Albert-Ludwigs-Universität Freiburg, Freiburg im Breisgau

Introduction

The classification of individuals into ancestral populations, i.e., inferring biogeographical ancestry (BGA) from DNA traces, is an important task in forensic genetics. Tools such as SNIPPER or the Admixture Model are widely used in forensic practice. However, they have not yet been systematically compared to classifiers widely used in other disciplines. Moreover, while established methods provide strong intercontinental classification, prediction accuracy within Europe remains limited. A

major factor contributing to this limitation is the set of genetic markers used for classification. Recent advances in machine learning offer new opportunities to improve both ancestry classification and the selection of ancestry-informative markers (AIMs).

Methods

We investigated machine learning approaches for BGA prediction at continental and intra-continental levels using published genomic datasets. First, the tabular foundation model TabPFN was benchmarked against established forensic classifiers. Second, a machine learning-based marker selection pipeline using permutation importance and deep learning models was applied to identify optimized AIM panels for intra-European classification. Marker sets containing 50-200 markers were evaluated using repeated cross-validation and compared with the VISAGE Enhanced Tool and allele frequency-based approaches. Third, we developed a method that detects, if the ancestral population of the trace is represented in the training data set.

Results

TabPFN consistently achieved the best overall classification performance across accuracy, ROC AUC, and log loss. At the continental level, classification accuracy increased from 84% using SNIPPER to 93% using TabPFN. For intra-European classification, accuracy increased from 43% to 48% across ten populations. Optimized marker selection further improved intra-continental prediction. For four European populations, accuracy increased from 68.2% using 104 VISAGE-selected markers to 73.7% with 100 newly selected markers and to 82.3% with 200 markers.

Discussion

Our findings demonstrate that machine learning improves both classification performance and marker selection for forensic ancestry prediction. In particular, tailored marker panels markedly enhance intra-European classification. The combination of advanced classifiers and optimized ancestry-informative markers enables more precise differentiation between genetically similar populations and highlights the potential of machine learning for fine-scale forensic ancestry analysis across different geographic regions.

[01-P1-21] Eliciting associations between clinical variables from LLMs via comparison questions across populations

12:20-12:25

Presenter: Fabian Kabus

Authors: Fabian Kabus [1]; Kian Kordtomeikel [2]; Thomas Brox [2]; Heinz Wiendl [3]; Daiana Stolz [4]; Harald Binder [5]

Affiliations: [1] Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Germany, Freiburg im Breisgau; [2] Department of Computer Science, Faculty of Engineering, Freiburg im Breisgau; [3] Klinik für Neurologie und Neurophysiologie - Universitätsklinikum Freiburg, Freiburg; [4] Clinic of Pneumology, Medical Center - University of Freiburg, Freiburg im Breisgau; [5] University of Freiburg, Freiburg, Germany, Freiburg

Introduction

Large language models (LLMs) are increasingly relevant to clinical decision tasks, but it remains unclear how to reliably query them and formalize the clinical knowledge they represent. We focus on associations between patient characteristics, including correlations and causal links. Directly asking for the correlation between two markers in a population can retrieve memorized statements or produce ungrounded numbers. We therefore treat the LLM as an experimental subject rather than a self-reporting agent.

Methods

Inspired by cognitive-science work on comparison judgments, we probe the model with patient-comparison triplets. For a pair of clinical variables, we ask whether a reference patient is more similar to one of two comparison patients while varying values on a dense grid. We fit a logistic surrogate model to the resulting binary choices and convert the fitted decision surface into a pairwise correlation estimate, without access to model internals. We then synthesize subpopulations through prompt-level environment shifts, estimate environment-specific correlations, and apply invariant causal prediction (ICP) as a conservative filter for candidate parent links. We demonstrate the framework in chronic obstructive pulmonary disease (COPD) and multiple sclerosis (MS).

Results

The elicited decision surfaces are smooth and well fit by the surrogate model. The recovered correlations are clinically plausible. In COPD, the method recovers a strong negative association between airflow obstruction and air trapping or hyperinflation, alongside positive associations within spirometric and diffusion-based measurements. In MS, the associations align with known relationships among imaging, fluid biomarkers, and disability-related measures. Compared with direct numeric correlation queries, the triplet-based estimates are less noisy and more sensitive to population shifts. Across environments, this sensitivity is sufficient for ICP to yield a small set of conservative invariant links.

Discussion

These results suggest that structured comparison questions can recover association structure from LLMs more reliably than direct elicitation. Prompted environments then make it possible to compare these associations across subpopulations and to move from implicit correlations toward causal statements. The framework remains exploratory and limited to model-implied rather than empirical clinical knowledge, but it offers a practical route for using LLMs as a source of prior knowledge in biomedical research.

SmallData Symposium Session 3: Uncertainty

Thu, 01.10.26, 13:30–15:00, Room: Rolf Böhme Saal

Chair: Carola S. Heinzel

Session 3: Uncertainty.

13:30-13:50 Heinz Wiendl and Susanne Weber: Clinical categories to statistical trajectories: An interdisciplinary approach to modeling disease progression in multiple sclerosis.

13:50-14:10 Martin Wolkewitz and Maria Elena Maccari: Challenges of target trial emulation in rare pediatric diseases: from limited data to clinical management.

14:10-14:50 Panel Discussion: The role of uncertainty in small data. Panelists: Heinz Wiendl, Martin Wolkewitz, Maria Elena Maccari, Peter Pfaffelhuber.

14:50-15:00 Harald Binder: Outlook on the small data landscape.

From Clinical Categories to Statistical Trajectories: An Interdisciplinary Approach to Modeling Disease Progression in Multiple Sclerosis

13:30-13:50

Presenter: Heinz Wiendl

Authors: Susanne Weber [1]; Heinz Wiendl [2]

Affiliations: [1] Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center - University of Freiburg, Germany, Freiburg; [2] Klinik für Neurologie und Neurophysiologie - Universitätsklinikum Freiburg, Freiburg

Introduction

Clinical reasoning frequently relies on disease categories and stages, whereas statistical models often describe disease progression as a continuous process. Translating clinical questions into appropriate quantitative models therefore requires close interdisciplinary collaboration between clinicians and statisticians. Using multiple sclerosis (MS) as an example, we present a collaborative workflow for developing and interpreting longitudinal disease progression models from both clinical and statistical perspectives, with a focus on their implications for clinical routine and decision-making.

Methods

The project was developed jointly by clinicians and statisticians in an iterative process involving formulation of clinically relevant questions, assessment of data quality and measurement properties, model selection, and interpretation of latent disease trajectories. Longitudinal MS data included repeated follow-up visits with heterogeneous clinical markers comprising continuous, binary, and count variables. To obtain a common representation suitable for dynamic modeling, markers were summarized using factor analysis. Disease progression was then modeled using two complementary

approaches: a continuous latent state-space model estimated via Kalman filtering and smoothing, and a discrete hidden Markov model (HMM). Particular emphasis was placed on discussing assumptions, interpretability, and the implications of continuous versus categorical representations of disease progression.

Results

The interdisciplinary exchange substantially influenced both model development and interpretation. Clinical expertise guided interpretation of latent dimensions, and evaluation of biological plausibility, while statistical modeling highlighted the consequences of discretizing potentially continuous disease processes. The state-space model yielded smooth patient-specific trajectories within a shared latent disease space, whereas the HMM represented disease evolution through transitions between discrete latent states. Comparing both approaches facilitated discussion about how methodological assumptions shape clinical interpretation, communication of disease progression, and potentially treatment-related decision-making in routine care.

Discussion

This project illustrates how interdisciplinary collaboration can improve the development, interpretation, and communication of complex longitudinal models in clinical research. Beyond the specific MS application, the study highlights how modeling decisions influence everyday clinical practice, including patient monitoring, treatment escalation, and communication of disease status. Continuous dialogue between clinicians and statisticians is therefore essential not only for methodological rigor but also for ensuring that statistical models remain clinically interpretable, actionable, and relevant for routine care.

Challenges of target trial emulation in rare pediatric diseases: from limited data to clinical management

13:50-14:10

Presenter: Martin Wolkewitz, Maria Elena Maccari

Authors: Martin Wolkewitz

Affiliations: Institut für Med. Biometrie und Med. Informatik, Freiburg

Introduction

Clinical research in rare pediatric diseases with highly complex management settings is fundamentally limited by small, heterogeneous, and sparsely observed patient cohorts, often including only a few hundred individuals. Despite these limitations, such cohorts frequently represent the best and sometimes only available source of high-quality clinical evidence. Their limited size, irregular longitudinal follow-up, heterogeneous disease courses, and incomplete covariate assessment create substantial challenges for conventional causal inference and machine-learning methods, which are typically developed for large, well-characterized populations with dense and standardized follow-up.

Methods

We use two cohort studies of rare monogenic immunological disorders as illustrative examples. These disorders are characterized by complex disease trajectories and multiple therapeutic options, including hematopoietic stem-cell transplantation. The cohorts are used to highlight key methodological challenges, including limited statistical power, treatment heterogeneity, time-dependent clinical decisions, irregular data collection, and missing information. We discuss how causal modeling strategies and target trial emulation frameworks could be adapted to support clinically meaningful inference in rare pediatric diseases

Results

The two cohort studies illustrate that rare disease datasets, although limited in scale, can provide a valuable framework for addressing clinically relevant treatment questions when analyzed using appropriately adapted causal methods. Target trial emulation offers a structured approach to defining treatment strategies, eligibility criteria, follow-up, outcomes, and causal contrasts, thereby improving transparency and reducing bias in observational analyses. At the same time, expert knowledge is essential for specifying plausible causal structures, identifying relevant confounders, handling sparse longitudinal data, and interpreting findings in the context of complex clinical decision-making. These examples demonstrate that combining available cohort data with disease-specific expertise can strengthen the credibility and clinical relevance of evidence generated in rare pediatric diseases.

Discussion

Rare pediatric diseases require causal inference approaches that are tailored to small, heterogeneous, and incompletely observed cohorts. Adapted target trial emulation frameworks, supported by modern causal models and expert-informed assumptions, can help generate robust and clinically interpretable evidence in settings where randomized trials are often infeasible. The methodological principles discussed here are not limited to monogenic immunological disorders but are transferable to other rare disease contexts with complex longitudinal treatment pathways and limited data availability.

Poster sessions

Poster Session: P1 - SmallData Similarity

[01-P1-16] Dissecting the molecular basis of monogenic neurodevelopmental disorders in human

Presenter: Melanie Kurzmann

Authors: Melanie Kurzmann [1]; Adrian Fritz [2]; Rolf Backofen [3]; Miriam Schmidts [4]

Affiliations: [1] Universitätsklinikum Freiburg, Freiburg im Breisgau; [2] University of Freiburg, Freiburg; [3] University of FreiburgDepartment of Computer ScienceBioinformatics, Freiburg im Breisgau; [4] Universitätsklinikum Freiburg, Zentrum für Kinderheilkunde und Jugendmedizin, Freiburg im Breisgau

Introduction

Neurodevelopmental disorders, including intellectual disability (ID), affect 1-3% of the global population, posing substantial healthcare challenges. Among these, monogenic forms represent a significant etiological subset. Our research focuses on unraveling the genetic basis of ID and characterising its molecular mechanisms. We employ exome sequencing for gene discovery and utilise an AI-based neural network causal gene prioritisation. To date, we have identified two novel human disease genes, TTBK1 and FKBP4, associated with syndromic ID phenotypes resulting from loss-of-function mutations. While TTBK1 has been implicated in Alzheimer's disease through its role in Tau phosphorylation, its neurodevelopmental loss-of-function effects remain unexplored. FKBP4, on the other hand, is involved in protein folding and mTOR signaling, with Fkbp4 knockout in mice linked to disrupted androgen receptor signaling and sexual development abnormalities.

Methods

We generate Ttbk1- and Fkbp4 mutant neural cell models using CRISPR/Cas and future work aims at functional characterisation of those models, including neural organoids. Likewise, we plan to generate zebrafish models to confirm the in vitro results.

Results

The knockout models in the mouse neural stem-cell line NE-4C have been established for both genes.Characterisation of NE-4C models using neuronal differentiation and transcriptomic analyses is currently ongoing.

Discussion

We have identified TTBK1- and FKBP4-loss-of function as novel cause of monogenic syndromal ID. Developing and characterising neural cell models will enhance our understanding of the underlying molecular mechanisms, and potentially identify new therapeutic targets. The AI approach for gene

prioritisation in exome datasets will accelerate the identification of additional causative genes in neurodevelopmental disorders.

[01-P1-18] Testing for Single-Population Ancestry in the Admixture Model

Presenter: Zoe Kristin Lange

Authors: Holger Dette [1]; Carola S. Heinzel [2]; Zoe Kristin Lange [1]; Peter Pfaffelhuber [2]

Affiliations: [1] Faculty of Mathematics, Ruhr-University Bochum, Bochum; [2] Department of Mathematical Stochastics, University of Freiburg, Freiburg

Introduction

Ancestry analysis is an important task in population genetics and forensic genetics, where genetic marker data are used to study population structure, infer biogeographical ancestry, and support the interpretation of genetic traces. Hard classifiers can provide useful single-population assignments when one ancestral population is clearly dominant. In contrast, admixture models represent an individual's genome as a mixture of allele contributions from multiple ancestral populations and are more informative when multiple ancestral populations contribute meaningfully to the genome. In practice, however, it is often unclear which interpretation is justified for a given individual. We propose a statistical test to assess whether a single ancestry population explains an individual's genome sufficiently well.

Methods

In a supervised admixture model with given reference allele frequencies, we test whether the dominant ancestry component exceeds a practitioner-chosen threshold that encodes the scientific meaning of single-population ancestry. Because the number of available genetic markers is often limited, we develop a constrained parametric bootstrap procedure designed to provide reliable calibration in small-sample settings. We additionally propose a sensitivity analysis over a small grid of dominance thresholds to assess how robust the conclusion is with respect to the chosen definition of single-population ancestry.

Results

Under standard regularity conditions for maximum likelihood estimation, we establish asymptotic type I error control and consistency of the proposed test. Thus, the procedure controls the probability of incorrectly declaring single-population ancestry at a chosen significance level, while detecting dominant ancestry components with probability tending to one. Simulation studies demonstrate good finite-sample error control and high power across different thresholds and admixture scenarios. We further illustrate the practical utility of the method using real genetic data from the 1000 Genomes Project.

Discussion

The proposed method provides an interpretable, threshold-based assessment of single-population ancestry with rigorous error control and strong empirical performance when the number of genetic markers is limited. The theory allows marker-wise heterogeneity, since loci are independent but have different allele-frequency profiles. More broadly, the constrained bootstrap principle can be adapted to other threshold-based testing problems with independent non-identically distributed observations.

[01-P1-19] Convergence rate analysis of the nonparametric maximum likelihood estimator with external data

Presenter: Nils Kober

Authors: Nils Kober [1]; Max Behrens [2]; Harald Binder [2]; Angelika Rohde [1]

Affiliations: [1] Department of Mathematical Stochastics, University of Freiburg, Freiburg; [2] Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg

Introduction

Clinical prediction models commonly have to learn from small, heterogeneous biomedical datasets, while potentially useful external cohorts are collected following different sampling schemes or measurement conventions. In such settings, additional external data are valuable to the estimation on the target dataset only if their underlying distribution is similar in a suitable sense. We develop mathematical theory for the estimation of the cumulative distribution function of a random variable from current-status observations. To improve the accuracy for small target sample sizes in this classical model, we include observations from a larger external dataset with a similar but in general non-identical distribution.

Methods

We propose a weighted nonparametric maximum likelihood estimator that uses a convex combination of the two datasets in its criterion function. For desirable rates of convergence, the weights need to be chosen based on the Hellinger distance between the true underlying distributions of the observed data.

Results

We derive its rate of convergence in Hellinger distance, which exceeds the n^frac13-rate of the classical NPMLE for monotone functions if the external sample size grows faster than the target sample size and the external distribution converges suitably to the target distribution. Our main contribution is the derivation of maximal inequalities to bound the expected supremum of the weighted empirical process mathbbGnwn := wn 2-1 sumi=1n wn,i (deltaXi - mathbbPi) over a function class, where (Xi sim mathbbPi)i=1,dots, n are independent but not necessarily identically distributed and wn = (wn,1, dots, wn,n)^top are nonnegative, normalized weights.

Discussion

Currently, we only consider deterministic weights in the criterion function of the NPMLE. For real-world application of the method, they would need to be chosen adaptively based on observed data, which is an open research question so far. Our methodology is closely related to the work of Behrens et al., who developed a weighted parametric maximum likelihood estimator to fit local regression model for different patient subgroups. References [1] Kober, N. University of Freiburg. Master's thesis: Convergence rate analysis of the nonparametric maximum likelihood estimator with external data. Not published. Jan 2026. [2] Behrens M, Stolz D, Papakonstantinou E, Nolde JM, Bellerino G, Rohde A, Hess M, Binder H. Contrasting Global and Patient-Specific Regression Models via a Neural Network Representation, Biometrical J. Mar 2026.

[01-P1-21] Eliciting associations between clinical variables from LLMs via comparison questions across populations

Presenter: Fabian Kabus

Authors: Fabian Kabus [1]; Kian Kordtomeikel [2]; Thomas Brox [2]; Heinz Wiendl [3]; Daiana Stolz [4]; Harald Binder [5]

Affiliations: [1] Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Germany, Freiburg im Breisgau; [2] Department of Computer Science, Faculty of Engineering, Freiburg im Breisgau; [3] Klinik für Neurologie und Neurophysiologie - Universitätsklinikum Freiburg, Freiburg; [4] Clinic of Pneumology, Medical Center - University of Freiburg, Freiburg im Breisgau; [5] University of Freiburg, Freiburg, Germany, Freiburg

Introduction

Large language models (LLMs) are increasingly relevant to clinical decision tasks, but it remains unclear how to reliably query them and formalize the clinical knowledge they represent. We focus on associations between patient characteristics, including correlations and causal links. Directly asking for the correlation between two markers in a population can retrieve memorized statements or produce ungrounded numbers. We therefore treat the LLM as an experimental subject rather than a self-reporting agent.

Methods

Inspired by cognitive-science work on comparison judgments, we probe the model with patient-comparison triplets. For a pair of clinical variables, we ask whether a reference patient is more similar to one of two comparison patients while varying values on a dense grid. We fit a logistic surrogate model to the resulting binary choices and convert the fitted decision surface into a pairwise correlation estimate, without access to model internals. We then synthesize subpopulations through prompt-level environment shifts, estimate environment-specific correlations, and apply invariant causal prediction (ICP) as a conservative filter for candidate parent links. We demonstrate the framework in chronic obstructive pulmonary disease (COPD) and multiple sclerosis (MS).

Results

The elicited decision surfaces are smooth and well fit by the surrogate model. The recovered correlations are clinically plausible. In COPD, the method recovers a strong negative association between airflow obstruction and air trapping or hyperinflation, alongside positive associations within spirometric and diffusion-based measurements. In MS, the associations align with known relationships among imaging, fluid biomarkers, and disability-related measures. Compared with direct numeric correlation queries, the triplet-based estimates are less noisy and more sensitive to population shifts. Across environments, this sensitivity is sufficient for ICP to yield a small set of conservative invariant links.

Discussion

These results suggest that structured comparison questions can recover association structure from LLMs more reliably than direct elicitation. Prompted environments then make it possible to compare these associations across subpopulations and to move from implicit correlations toward causal statements. The framework remains exploratory and limited to model-implied rather than empirical clinical knowledge, but it offers a practical route for using LLMs as a source of prior knowledge in biomedical research.

[01-P1-22] A Task-First Specification Protocol for Similarity in Data Integration

Presenter: Max Behrens

Authors: Max Behrens [1]; Nils Kober [2]; Maryam Farhadizadeh [3]; Hannah Habenicht [4]; Maren Hackenberg [5]; Fabian Kabus [6]; Michelle Pfaffenlehner [7]; Clemens Schächter [8]; Harald Binder [9]; Nadine Binder [10]; Johannes Hertel [11]

Affiliations: [1] Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, Freiburg; [2] University of Freiburg, Freiburg; [3] Institute of General Practice/Family Medicine, Medical Faculty and Medical Center, University of Freiburg, Freiburg im Breisgau; [4] Department of Biometry and Environmental System Analysis, University of Freiburg, Freiburg; [5] European Molecular Biology Laboratory, Heidelberg; [6] Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Germany, Freiburg im Breisgau; [7] Institut für medizinische Biometrie und Statistik, Universität Freiburg, Freiburg; [8] Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center - University of Freiburg, Germany, Freiburg; [9] University of Freiburg, Freiburg, Germany, Freiburg; [10] Medical Center - University of Freiburg, Freiburg im Breisgau; [11] University Medicine Greifswald, GREIFSWALD

Introduction

There are various tasks, where researchers need to integrate data from multiple sources, when wanting to extend the population a model covers, to combine evidence across study types, to replicate findings, or to enhance statistical power when individual studies are limited. Implicitly, each task relies on assumptions of similarity between data sources. If analysis workflows consider quantifying this uncertainty at all, the danger is to just assess types of similarity that are convenient to implement, so issues from a lack of similarity might be overlooked.

Methods

We introduce the similarity chain, a protocol where the task at hand steers similarity assessment. It relies on four interlinked components: the analytical task (the integration goal), the relevant type of similarity (data, design, or model), a plain-language condition (what similarity must satisfy), and an empirical check (how to assess it). Data similarity refers to the observations themselves, design similarity to how they were collected, and model similarity to the analytical model applied. The protocol also highlights critical assumptions. If these are not fulfilled, subsequent checks are invalidated and integration may not be recommended.

Results

We illustrate the protocol across four worked examples in cohort pooling, prediction model transport, evidence synthesis between RCTs and real-world data, and cross-species transfer in single-cell genomics. In each case, the similarity chain flagged assumptions that conventional checks miss and recommended whether integration is reasonable. When transporting a cardiovascular risk score to a different population, for example, preserved discrimination might have been mis-read as evidence of transportability and the protocol flagged that the score can miscalibrate when the baseline event rate differs.

Discussion

Thus, the similarity chain provides a structured way to reason about when integration is feasible and what to report when it is not. Alongside, we provide a checklist mapping analytical tasks to their similarity requirements, reference catalogs of conditions and empirical checks, and a conversational AI skill that elicits the task, proposes the relevant similarity type, suggests empirical checks, and drafts the report. Given that no single validation establishes a model as fit for all settings, the protocol turns similarity assumptions into reportable artifacts for transparent, reproducible integration decisions.

[01-P1-24] VAEs and nonlinear filtering for SMA registry data

Presenter: Felix Barrez Tambe Ndonfack

Authors: Felix Barrez Tambe Ndonfack [1]; Clemens Schächter [2, 3]; Harald Binder [2, 3, 4]; Thorsten Schmidt [1]

Affiliations: [1] Department of Mathematical Stochastics, Mathematical Institute, University of Freiburg, Freiburg; [2] Institute of Medical Biometry and Statistics (IMBI), Medical Faculty and Medical Center – University of Freiburg, Germany, Freiburg; [3] Freiburg Center for Data Analysis, Modeling and AI – University of Freiburg, Germany, Freiburg; [4] CIBSS, Centre for Integrative Biological Signalling Studies – University of Freiburg, Germany, Freiburg

Introduction

Rare disease registries, such as those in spinal muscular atrophy (SMA), present a distinctive statistical challenge: patients are observed at scheduled clinical visits, treatment switches occur at

predictable times, and motor function evolves continuously between observations. Standard mixed-effects models require explicit feature engineering to represent known event times and typically do not model latent continuous-time dynamics between visits. We need a framework that models signal dynamics between visits, accounts for scheduled treatment-switch discontinuities, and provides calibrated uncertainty at the individual patient level.

Methods

We apply the nonlinear filtering framework with stochastic discontinuities of [1] to real SMA registry data. The unobserved patient health state is modelled as a multivariate Ornstein-Uhlenbeck process in three motor dimensions (sitting, standing, walking), observed at predictable visit times with additive noise. Treatment switches enter the model as predictable jump times, at which the filter covariance is inflated via a Q-switch term. A LLM is fitted in parallel as a population-level comparator, and a synthetic ground truth experiment with injected score jumps validates filter sensitivity.

Results

Applying the 3D Kalman filter to real SMA registry data, filter innovations at treatment-switch visits are significantly larger than at regular visits across all three motor dimensions. The effect is dimension-specific, with sitting most affected and walking showing a floor effect. A multivariate finding i.e. the standing-walking innovation correlation decreases at switch visits, is invisible in the scalar sum score. The LMM confirms significant post-switch score improvement across dimensions, consistent with the filter results. A synthetic detection experiment characterises filter sensitivity and shows that controlling the false positive rate requires a data-driven threshold more conservative than the standard 2-sigma rule.

Discussion

This work provides an empirical validation of the stochastic discontinuity filtering framework on real clinical data, offering a principled approach to treatment switch detection in rare disease registries without requiring a uniform effect direction across patients. The 3D extension reveals dimension-specific treatment responses masked by the scalar sum score. Limitations include moderate filter undercoverage, attributable to heavy-tailed score variability not fully captured by the OU model. Reference [1] Thorsten Schmidt and Félix B. Tambe-Ndonfack (2026), Nonlinear filtering with stochastic discontinuities. Preprint arXiv:2605.12142

[01-P1-25] Using latent representations to link disjoint longitudinal data for mixed-effects regression

Presenter: Clemens Schächter

Authors: Clemens Schächter [1, 2]; Maren Hackenberg [1, 2]; Michelle Pfaffenlehner [1, 2]; Felix Barrez Tambe Ndonfack [3]; Thorsten Schmidt [3]; Astrid Pechmann [4]; Janbernd Kirschner [4]; Jan Hasenauer [5, 6]; Harald Binder [1, 2, 7]

Affiliations: [1] Institute of Medical Biometry and Statistics (IMBI), Medical Faculty and Medical Center, University of Freiburg, Freiburg; [2] Freiburg Center for Data Analysis, Modeling and AI, University of Freiburg, Freiburg; [3] Department of Mathematical Stochastics, Mathematical Institute, University of Freiburg, Freiburg; [4] Department of Neuropediatrics and Muscle Disorders, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg im Breisgau; [5] Bonn Center for Mathematical Life Sciences, University of Bonn, Bonn; [6] Life and Medical Sciences (LIMES) Institute, University of Bonn, Bonn; [7] CIBSS, Centre for Integrative Biological Signalling Studies, University of Freiburg, Freiburg

Introduction

Many rare diseases offer limited established treatment options, leading patients to switch therapies when new medications emerge. To analyze the impact of such treatment switches within the low sample size limitations of rare disease trials, it is important to use all available data sources. This, however, is complicated when the use of measurement instruments changes during the observation period, for example when instruments are adapted to specific age ranges. The resulting disjoint longitudinal data trajectories complicate the application of traditional modeling approaches like mixed-effects regression.

Methods

We tackle this by mapping observations of each instrument to an aligned low-dimensional temporal trajectory, enabling longitudinal modeling across instruments. Specifically, we employ a set of variational autoencoder architectures to embed item values into a shared latent space for each time point. Temporal disease dynamics and treatment switch effects are then captured through a mixed-effects regression model applied to latent representations. To enable statistical inference, we present a novel statistical testing approach that accounts for the joint parameter estimation of mixed-effects regression and variational autoencoders.

Results

The methodology is applied to quantify the impact of treatment switches for patients with spinal muscular atrophy. Here, our approach aligns motor performance items from different measurement instruments for mixed-effects regression and maps estimated effects back to the observed item level to quantify the treatment switch effect. Our approach allows for model selection as well as for assessing effects of treatment switching.

Discussion

The results highlight the potential of modeling in joint latent representations for addressing small data challenges in rare disease research. By combining neural-network-based measurement alignment with longitudinal mixed-effects regression, the proposed framework enables the use of heterogeneous item-level data while retaining statistical inference. This is particularly relevant when changes in measurement instruments would otherwise restrict analyses to smaller and potentially biased subsets, although careful validation remains necessary for interpreting treatment-switch effects.

[01-P1-27] CRISCross: Modeling Sequence-Context Interactions for Robust CRISPR-Cas9 Off-Target Prediction Across Cell Types

Presenter: William Jobson Pargeter

Authors: William Jobson Pargeter [1]; Rolf Backofen [1]; Toni Cathomen [2]; Masako Kaufmann [2]

Affiliations: [1] University of Freiburg Department of Computer Science Bioinformatics, Freiburg; [2] Institute for Transfusion Medicine and Gene Therapy, University Medical Center Freiburg, Freiburg

Introduction

CRISPR-Cas9 enables precise genome editing but remains limited by off-target effects, raising safety concerns and necessitating accurate prediction. Although epigenetic features influence Cas9 binding and cleavage, their predictive signals are often weak, highly variable, and context-dependent, limiting the effectiveness of existing machine learning approaches, particularly in cross-cell-type settings where epigenetic landscapes differ substantially, and models trained on one cell type frequently fail to generalize. Here, we systematically investigate the role of epigenetic context in off-target prediction and demonstrate that its utility depends on modeling interactions with sequence features rather than treating it as an independent signal. Tabular and deep learning experiments show that epigenetic features provide a consistent but modest complementary signal, and that late-stage fusion strategies are insufficient to exploit this information effectively.

Methods

To address this, we introduce CRISCross, cross-attention-based architecture that captures sequence-epigenetic interactions at the nucleotide level, alongside CRISPRAT, a synthetic pretraining strategy that learns transferable representations from genome-derived sgRNA-off-target pairs without relying on dataset-specific pretraining.

Results

CRISCross, incorporating AlphaGenome-derived chromatin accessibility signals, achieves an AUC-PR of 0.448 in the cross-cell-type setting, compared to 0.345 for the best-performing existing model, demonstrating improved generalization across guide RNAs, datasets, and cellular contexts.

Discussion

Together, these results show that accurate CRISPR off-target prediction requires interaction-aware integration of epigenetic context alongside generalizable pretraining strategies.

[01-P1-28] Benchmarking single-cell large language models

Presenter: Adrian Fritz

Authors: Adrian Fritz [1]; Niklas Brunn [2]; Jonatan Menger [3]; Anup Kumar [1]; Sonia Maria Krißmer [4]; Pavankumar Videm [1]; Ryan Köksal [1]; Maren Hackenberg [5]; Rolf Backofen [1]

Affiliations: [1] Department of Computer Science, Faculty of Engineering, University of Freiburg, Freiburg; [2] Institute of Medical Biometry and Statistics (IMBI), Faculty of Medicine and Medical Center, University of Freiburg, Germany, Freiburg; [3] Institute of Medical Biometry and Statistics (IMBI), Faculty of Medicine and Medical Center, University of Freiburg, Freiburg; [4] Medical Center Freiburg, Freiburg; [5] Institute of Medical Biometry and Statistics, University of Freiburg, Freiburg

Introduction

Foundation models are rapidly expanding in single-cell omics, but it remains unclear how language-based approaches compare with transcriptome-native models for biological interpretation. Recent work spans prompted general-purpose large language models, models that treat transcriptomic data as language, multimodal models combining text and quantitative information and expression-focused single-cell foundation models, yet these approaches are often evaluated in separate settings and under different assumptions.

Methods

We investigate this varying landscape through a comparative benchmarking framework with a particular focus on language-based and language-integrated foundation models for single-cell analysis. We center the comparison on cell type annotation as a practical and widely used benchmark task, while recognizing that cellular identity is often continuous and context-dependent rather than fully captured by discrete labels. Within a modular evaluation workflow that standardizes dataset specification, model wrapping, ontology-aware label harmonization and metric aggregation, we compare prompted generalist large language models, transcriptome-to-language models and multimodal text-cell models against expression-native single-cell foundation models and conventional annotation baselines. The benchmark is completely reproducible and publicly available as a nextflow workflow including both the model implementations and data sets. The modular setup facilitates easy extension to further data sets and models, not only for benchmarking purposes but even for usage on new data sets or for ensemble predictions.

Results

Across benchmark tasks, language-based and language-integrated models achieved reasonable annotation performance but did not consistently outperform transcriptome-native foundation models or established annotation approaches. In addition, prompted language models occasionally produced malformed, non-biological or contextually inconsistent labels, highlighting current robustness limitations in open-ended annotation settings.

Discussion

These findings suggest that language-oriented single-cell models are promising for flexible and multimodal biological interpretation, but currently complement rather than replace established transcriptome-centric methods. The presented benchmarking framework provides a reproducible basis for future evaluation of robustness, interpretability and open-ended language-driven analysis tasks in single-cell omics.

[01-P1-30] A Sequence Alignment-Based Approach to Identifying Subgroups in Longitudinal Healthcare Data

Presenter: Maryam Farhadizadeh

Authors: Maryam Farhadizadeh [1]; Maria Weymann [2]; Harald Binder [2]; Daiana Stolz [3]; Jan Hasenauer [4]; Nadine Binder [1]

Affiliations: [1] Institute of General Practice/Family Medicine, Medical Faculty and Medical Center, University of Freiburg, Freiburg im Breisgau; [2] Institute of Medical Biometry and Statistics (IMBI), Medical Faculty and Medical Center, University of Freiburg, Freiburg im Breisgau; [3] Clinic of Pneumology, Medical Center - University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany, Freiburg im Breisgau; [4] Bonn Center for Mathematical Life Sciences, University of Bonn, Bonn

Introduction

Identifying patient subgroups from longitudinal healthcare pathways is challenging because patients rarely follow identical sequences of care. Patient trajectories differ in the ordering, length, and composition of healthcare events, while clinically similar events may be represented using different medical codes. Consequently, current approaches often rely on exact matching of clinical codes and therefore fail to capture semantic relationships between clinically related healthcare events represented using different medical codes, potentially overlooking meaningful similarities between patient trajectories. Furthermore, many existing methods focus only on clustering patient trajectories and do not identify typical healthcare pathways within the resulting patient subgroups.

Methods

To identify patient subgroups from longitudinal healthcare data, similarities between healthcare pathways were quantified using pairwise sequence alignment. Because clinically related healthcare events may be represented using different medical codes, direct comparison of raw codes may fail to capture semantic similarity between clinical events. To address this, the textual descriptions of clinical codes were embedded using a domain-specific transformer language model to capture relationships between medical concepts. These embeddings were then used to construct a substitution matrix for sequence alignment, such that semantically related events received lower mismatch penalties than clinically unrelated events, while events from different clinical categories were considered less interchangeable. The resulting alignment-based similarities were used for hierarchical clustering to identify patient subgroups. Within each subgroup, multiple sequence alignment was performed using MAFFT to identify typical healthcare pathways.

Results

The proposed framework was applied to longitudinal routine care data from patients with chronic obstructive pulmonary disease (COPD) treated at the University Medical Center Freiburg. The dataset contained heterogeneous healthcare pathways characterized by diverse comorbidity patterns, repeated hospitalizations, and varying clinical event sequences. Alignment-based clustering identified clinically interpretable patient subgroups with distinct combinations of comorbidities and healthcare

trajectories. Multiple sequence alignment within clusters revealed distinct typical healthcare pathways and clinically meaningful differences in care trajectories across patient subgroups.

Discussion

By combining alignment-based clustering with multiple sequence alignment, the proposed framework identifies clinically interpretable patient subgroups together with their typical healthcare pathways. Incorporating semantic similarity between clinical events improves comparison of heterogeneous patient trajectories and interpretation of subgroup-

---

Poster Session: P2 - SmallData Transfer

[01-P2-01] Systematic Evaluation of Few-Shot Learning for Clinical RDEB Assessment

Presenter: Julia Hindel

Authors: Julia Hindel [1]; Jelena Bratulić [2]; Ayush Gupta [1]; Meropi Karakioulaki [3]; Cristina Has [3]; Thomas Brox [2]; Abhinav Valada [4]

Affiliations: [1] Department of Computer Science, Faculty of Engineering, University of Freiburg, Freiburg im Breisgau; [2] Computer Vision Group, University of Freiburg, Freiburg im Breisgau; [3] Department of Dermatology, Medical Faculty and Medical Center - University of Freiburg, Freiburg; [4] Department of Computer Science, University of Freiburg, Freiburg im Breisgau

Introduction

Recessive Dystrophic Epidermolysis Bullosa (RDEB) is a severe, inherited skin disorder characterized by chronic wounds, scarring, and a markedly increased risk of squamous cell carcinoma. Progress in automated clinical characterization and research is limited by scarce public image datasets and small patient cohorts, making supervised machine learning approaches infeasible due to their demand for large labeled datasets.

Methods

We present a case study leveraging a dataset of approximately 90 wound photographs from RDEB patients. To address data scarcity, we evaluate various few-shot learning techniques, including In-Context Learning, Inductive Few-Shot Learning, and Transductive Few-Shot Learning, to characterize wounds by cancer status, collagen levels, patient age, and anatomical location. We also benchmark several general and medical vision foundation models, assessing their baseline performance and examining the impact of wound area masking on model accuracy.

Results

Preliminary results indicate that unsupervised clustering on both general and specialized medical vision foundation models perform poorly in distinguishing the relevant categories with the given dataset. Through systematic evaluation, we find that cancer status and wound location can be reliably classified from clinical photographs using few-shot learning methods. In contrast, accurate characterization of patient age and collagen levels remains challenging and largely unresolved. Further, we identify that masking the wound region is essential for successful differentiation of cancer status while hindering location differentiation.

Discussion

Our findings highlight the current limitations of off-the-shelf vision foundation models for rare diseases such as RDEB under genuine clinical data scarcity. While few-shot learning methods offer some promise, particularly for identifying cancer and location, the generalizability and accuracy for other clinically relevant features, such as age and collagen content, are limited. This work underscores the need for tailored vision methods and further curation of rare disease datasets, contributing a systematic evaluation for RDEB and informing future approaches to medical image analysis in ultra-rare conditions.

[01-P2-03] Learning to Read Where to Look: Disease-Aware Vision-Language Pretraining for 3D CT

Presenter: Sebastian Walter

Authors: Simon Ging [1, 2]; Philipp Arnold [3]; Sebastian Walter [4]; Hani Alnahas [1]; Hannah Bast [5]; Elmar Kotter [3]; Jiancheng Yang [6, 7]; Behzad Bozorgtabar [2]; Thomas Brox [1]

Affiliations: [1] Computer Vision Group, University of Freiburg, Freiburg im Breisgau; [2] Adaptive & Agentic AI (A3) Lab, Aarhus University, Aarhus; [3] Department of Radiology, Medical Center - University of Freiburg, Freiburg im Breisgau; [4] Chair of Algorithms and Data Structures, University of Freiburg, Freiburg im Breisgau; [5] Department of Computer Science, University of Freiburg, Germany, Freiburg im Breisgau; [6] ELLIS Institute Finland, Espoo; [7] School of Electrical Engineering, Aalto University, Espoo

Introduction

Recent 3D CT vision-language models align volumes with reports via contrastive pretraining, but typically rely on limited public data and provide only coarse global supervision.

Methods

We train a 3D CT vision-language model on 98k report-volume pairs (50k patients) collected at a single hospital, combined with public datasets, using SigLIP-style contrastive pretraining together with prompt-based disease supervision in the shared vision-text embedding space.

Results

On CT-RATE, our model achieves state-of-the-art text-to-image retrieval (R@10 31.5 vs. 22.2) and competitive disease classification (AUC 83.8 vs. 83.8), with consistent results on Rad-ChestCT (AUC 77.0 vs. 77.3).

Discussion

We further observe that radiologists routinely reference specific images within their reports (e.g., "series X, image Y"), linking textual descriptions to precise axial locations. We automatically mine 262k such snippet-slice pairs and introduce the task of intra-scan snippet localization – predicting the axial depth referred to by a text snippet – reducing mean absolute error to 36.3 mm at 12 mm feature resolution, compared with 67.0 mm for the best baseline. Adding this localization objective leaves retrieval and classification broadly unchanged within confidence bounds, yielding a single unified model for retrieval, classification, and intra-scan grounding.

[01-P2-04] From Playground to Practice, building foundation models in hours

Presenter: Samuel Böhm, Johannes Hog

Authors: Samuel Böhm [1]; Johannes Hog [2]; Alexander Pfefferle [2]; Lennart Purucker [2]; Pascal Schlosser [3]; Frank Hutter [2]

Affiliations: [1] Institute of Epidemiology and Prevention, Faculty of Medicine and Medical Center - University of Freiburg, Freiburg im Breisgau; [2] Institut für Informatik, Albert-Ludwigs-Universität Freiburg, Freiburg im Breisgau; [3] Institute of Epidemiology and Prevention, Faculty of Medicine and Medical Center - University of Freiburg, Freiburg

Introduction

Tabular foundation models (TFMs) such as TabPFN have recently achieved strong performance on small- to medium-sized tabular prediction tasks across medicine, finance, and marketing. Inspired by these successes, the community has begun extending TFMs to new domains and downstream tasks. However, existing implementations suffer from convoluted architectures that are difficult to understand, and cumbersome to adapt thus slowing methodological progress and limiting accessibility for applied researchers.

Methods

We introduce TFMPlayground, a lightweight, modular, and well-documented codebase that lowers the barrier to entry for TFM research. TFMPlayground decomposes the core ingredients of TFM training, from prior sampling over architectures to evaluation pipelines, into small, self-contained components that can be independently modified and freely recombined with minimal friction. To demonstrate the framework's utility, we used it to build SurvPFN, a domain-specific TFM for survival analysis. Survival prediction from right-censored time-to-event data is central to clinical and

epidemiological research, where small cohort sizes are the norm. Leveraging TFMPlayground's modular design, we construct survival-specific priors based on structural causal models with parametric event-time distributions and censoring mechanisms, and train a survival-aware TFM end-to-end.

Results

On a collection of 22 real-world survival tasks, SurvPFN is competitive with classical and deep survival baselines, while requiring no per-dataset fitting, survival-specific architecture, or feature engineering, making it immediately deployable to unseen cohorts.

Discussion

SurvPFN illustrates how TFMPlayground enables rapid prototyping of new TFM variants, making it straightforward to move from domain insight to a working model. Central to this is the lightweight and fast-to-train TFMPlayground backbone that makes iterative design and systematic ablation studies genuinely feasible. We could train, evaluate, and discard candidate prototypes within a single working day, a feedback loop that would be prohibitively slow with heavier architectures. For SurvPFN, this meant systematically exploring priors, loss functions, and architectural choices. By offering a transparent and flexible foundation for TFM research, TFMPlayground broadens participation in tabular foundation modeling and encourages the search for creative solutions.

[01-P2-06] Process-guided neural networks for forest carbon flux modelling under water stress: implications for drought-occurrences under climate change

Presenter: Hannah Habenicht

Authors: Hannah Habenicht [1]; Hanne Raum [2]; Fabian Kabus [3]; Harald Binder [4]; Joschka Boedecker [2, 5]; Carsten F. Dormann [6]

Affiliations: [1] Department of Biometry and Environmental System Analysis, University of Freiburg, Freiburg im Breisgau; [2] Neurorobotics Lab, University Freiburg, Freiburg im Breisgau; [3] Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Germany, Freiburg im Breisgau; [4] University of Freiburg, Freiburg, Germany, Freiburg; [5] BrainLinks-BrainTools, University Freiburg, Freiburg im Breisgau; [6] Department of Biometry and Environmental System Analysis, University Freiburg, Freiburg im Breisgau

Introduction

Forests account for the largest share of terrestrial carbon uptake, yet the models we rely on to predict their carbon and water fluxes are overwhelmingly trained on humid, radiation-limited conditions. Under projected climate change, many forest regions will likely experience increased intensified seasonal drought making the capacity to transfer these models from radiation-limited to water-limited regimes not merely a methodological question, but a prerequisite for accurate carbon cycle projections.

Methods

We test this capacity directly. Using FLUXNET flux tower observations stratified by Köppen climate categorization, we train process-guided neural networks (PGNNs) exclusively on well-watered forest sites and evaluate transfer to summer-dry regimes, where we treat the spatial distributional shift as a tractable proxy for the hydrological transitions projected under changing climatic conditions. The process backbone is BASFOR, chosen for its explicit dynamic coupling between soil water status and carbon assimilation and atmospheric demand.

Results

Across summer-dry test sites, BASFOR performs very poorly as a standalone predictor under these out-of-distribution conditions. A purely data-driven model partially recovers predictive skill and improves a little further when trained on drought occurrences. The process-guided hybrid outperforms both, without requiring the drought occurrences training data.

Discussion

These results demonstrate that mechanistic water-stress coupling substitutes for information that would otherwise need to be measured directly, improving transfer to water-stressed regimes even when the process model itself generalises poorly. The benefit is conditional on the process backbone adequately representing the dominant limiting factor at the target site, but where that condition holds, hybrid process guidance offers a practical path towards more robust carbon flux prediction under climatically novel conditions.

[01-P2-07] Generating realistic bacterial microscopy from mechanistic simulators with limited real-world data

Presenter: Jelena Bratulić

Authors: Jelena Bratulić [1]; Moritz Steinmaier [1]; Jonas Pleyer [2]; Simon Ging [3]; Christian Fleck [2]; Thomas Brox [1]

Affiliations: [1] Computer Vision Group, University of Freiburg, Freiburg im Breisgau; [2] Freiburg Center for Data Analysis, Modeling and AI, University of Freiburg, Freiburg im Breisgau; [3] Computer Vision Group, University of Freiburg, Freiburg

Introduction

Microscopy-based downstream tasks such as cell segmentation and tracking rely on precisely annotated data, which is costly to obtain: experts must delineate cell boundaries by hand, and labeling lineages across cell divisions in video data is both laborious and inherently ambiguous. Agent-based simulators built on mechanistic cell and bacteria models, such as cr_mech_coli, bypass this annotation bottleneck by producing image-annotation pairs with pixel-perfect ground-truth masks, including cell positions, orientations, growth phases, and accurate lineage trees, under physically plausible growth dynamics. However, naïve renders of simulator output appear overly synthetic for

deep learning training, limiting generalization to real microscopy and constraining the practical use for large-scale synthetic data generation.

Methods

We build on the cr_mech_coli simulator for rod-shaped bacteria and present two approaches for closing the sim-to-real gap: a physics-inspired pipeline, crm_gen, and a learned style-transfer model, diff_gen. crm_gen renders synthetic micrographs across three modalities, bright-field, phase-contrast, and fluorescence, by composing independent modules that approximate sample absorption, optical point-spread-function blur, sensor noise, uneven illumination, and background structure. Custom modality styles are supported by automatically calibrating the modality-relevant module parameters via derivative-free optimization on a small reference set. diff_gen leverages a pretrained diffusion model fine-tuned with LoRA on a small reference set and a modality vector, conditioned either on the segmentation mask (style transfer only) or directly on the simulator's mechanistic state, replacing the rendering step.

Results

We evaluate both pipelines across all three modalities using image-similarity scores, distribution-shift measures, and qualitative comparison, and we analyze data robustness by varying the number of reference samples used for calibration and style transfer. Initial results indicate effective style transfer for both pipelines, even from small reference subsets.

Discussion

Classical calibration and learned diffusion-based style transfer close the sim-to-real gap through complementary mechanisms: crm_gen is fast and interpretable but requires per-style fitting, while diff_gen enables multi-style generalization at a higher resource cost. Although the gap is not closed in absolute terms, both pipelines demonstrate the value of agent-based simulators for generating large-scale synthetic data with accurate annotations, mitigating the need for expert labeling and supporting downstream training in small-data scenarios with limited real-world annotated data.

[01-P2-09] Using Knowledge Graphs to harvest datasets for efficient CLIP model training

Presenter: Sebastian Walter

Authors: Simon Ging [1]; Sebastian Walter [2]; Jelena Bratulić [3]; Johannes Dienert [3]; Hannah Bast [3]; Thomas Brox [4]

Affiliations: [1] University of Freiburg, Freiburg; [2] University of Freiburg, Freiburg im Breisgau; [3] Computer vision group, University of Freiburg, Freiburg im Breisgau; [4] Computer Vision Group, University of Freiburg, Freiburg im Breisgau

Introduction

Training high-quality CLIP models typically requires enormous datasets, which limits the development of domain-specific models – especially in areas that even the largest CLIP models do not

cover well – and drives up training costs. This poses challenges for scientific research that needs fine-grained control over the training procedure of CLIP models.

Methods

In this work, we show that by employing smart web search strategies enhanced with knowledge graphs, a robust CLIP model can be trained from scratch with considerably less data.

Results

Specifically, we demonstrate that an expert foundation model for living organisms can be built using just 10M images.

Discussion

Moreover, we introduce EntityNet, a dataset comprising 33M images paired with 46M text descriptions, which enables the training of a generic CLIP model in significantly reduced time.

[01-P2-10] Reinforcement Learning-Guided Surrogate Modeling for Efficient Emulation of Forest Process Models

Presenter: Hanne Raum, M Asif Hasan

Authors: Hanne Raum [1]; M Asif Hasan [2]; Hannah Habenicht [3]; André Biedenkapp [4]; Noor Awad [4]; Carsten F. Dormann [5]; Joschka Boedecker [6]

Affiliations: [1] Neurorobotics Lab, University Freiburg, Freiburg; [2] Machine Learning Lab, University of Freiburg, Freiburg; [3] Department of Biometry and Environmental System Analysis, University of Freiburg, Freiburg; [4] Machine Learning Lab, University of Freiburg, Freiburg im Breisgau; [5] Department of Biometry and Environmental System Analysis, University Freiburg, Freiburg im Breisgau; [6] Neurorobotics Lab, University Freiburg, Freiburg im Breisgau

Introduction

Environmental process models such as Preles typically require site-specific parameter calibration based on forest characteristics and local climatic conditions, which is computationally expensive and limits transferability across sites.

Methods

Adaptive surrogate models offer a promising alternative as they can more easily generalize to new sites and improve interoperability without requiring extensive retraining. In particular, our goal is to replace traditional process-based models with pretrained neural networks trained on a diverse range of inputs and parameter configurations. Such training data must cover a broad space of calibrated parameters and the corresponding outputs generated by models such as Preles. To achieve this, we first perform a global sensitivity analysis to identify the subset of environmental process model parameters that most strongly influence model output, reducing the effective parameter dimensionality. We then efficiently sample this reduced parameter space using Reinforcement Learning (RL),

where an agent learns a sampling policy that maximizes the informativeness of the collected (parameters, driver, output) tuples relative to those already gathered, ensuring broad and diverse coverage of the process model behavioral repertoire.

Results

Once a sufficiently diverse data set of parameter-output pairs has been generated, a neural network-based simulator can be trained on these synthetic data, using both model parameters and environmental drivers as input. The resulting simulator can then be adapted to real-world measurements, enabling efficient transfer to new forest sites with minimal additional calibration.

Discussion

A possible limitation of this approach is that the quality of the learned surrogate strongly depends on the coverage and representativeness of the sampled parameter-output space, which may not fully capture rare or extreme ecosystem dynamics. Future work will focus on assessing the robustness of the proposed framework under different environmental conditions.

[01-P2-12] Dual-Scale Metabolic Modeling Unveils Spatial Compartmentalization and Sexual Dimorphism in the Mouse Proximal Tubule

Presenter: Chenglong Huang, Daniel Fässler

Authors: Chenglong Huang [1]; Daniel Fässler [1]; Anna Köttgen [2]; Johannes Hertel [3]

Affiliations: [1] University Medicine Greifswald, Greifswald; [2] Institute of Epidemiology and Prevention, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg im Breisgau; [3] University Medicine Greifswald, GREIFSWALD

Introduction

The kidney comprises diverse cell populations with distinct metabolic functions that cannot be fully captured by conventional bulk or tissue-level metabolic reconstructions. This limitation is particularly pronounced in the proximal tubule (PT), where recent single-cell transcriptomic studies have revealed robust sex-specific and spatial metabolic specialization. To bridge this resolution gap, we developed a dual-scale metabolic network modeling framework for the mouse PT.

Methods

By integrating bulk tissue transcriptomes and single-cell RNA-sequencing data with a mouse genome-scale metabolic reconstruction, we constructed sex- and segment-specific (PTS1, PTS2, and PTS3) models at both the tissue level and single-cell resolution. Utilizing the enhanced Flux Potential Analysis (eFPA) method to constrain metabolic fluxes via transcriptomic penalties, we generated 20 bulk-tissue models alongside 501 single-cell bootstrap-derived models.

Results

All PT segments exhibit glycogenic capacity that exceeded their own glucose consumption potential. Beta-oxidation potential of both essential and non-essential fatty acids was substantial, while de novo fatty acid synthesis rates were low. Amino acid uptake was prevalent in all three segments with decreasing tendency across the segments, while uptake of small polypeptide was placed in the S2 and S3 segments in line with the literature. Both bulk and single-cell models consistently revealed a spatial gradient, with the activities of amino acid transport, the tricarboxylic acid (TCA) cycle, folate metabolism, the pentose phosphate pathway (PPP), and tetrahydrobiopterin progressively decreasing from the S1 to the S3 segment in a partly sex-specific way, hinting at metabolic vulnerabilities particular in male PTS3 cells. Furthermore, systematic knockout modelling revealed sex- and segment-specific metabolic effects of gene dysfunction in silico.

Discussion

Our study maps a high-resolution spatial metabolic zonation in the PT, and enables the investigation of the interplay between spatial organization and sex-specific metabolic flexibility in the proximal tubule in both physiological and pathophysiological states. Application cases could be in silico screening on drug targets and the interrogation of gene functions in the PT.

[01-P2-13] Latent Trajectory Identifiability with Neural ODEs in Small Data Regimes

Presenter: Timo Rachel

Authors: Timo Rachel [1]; Tim Litwin [1]; Hanning Yang [1]; Meropi Karakioulaki [2]; Cristina Has [2]; Moritz Hess [1]

Affiliations: [1] Institute of Medical Biometry and Statistics, Medical Center, Faculty of Medicine, University of Freiburg, Germany, Freiburg; [2] Department of Dermatology, Medical Faculty and Medical Center - University of Freiburg, Freiburg

Introduction

Inflammatory biomarkers such as C-reactive protein (CRP) and interleukin-6 (IL-6) provide crucial insight into disease activity and treatment response. In inflammatory diseases where these biomarkers play a central role, clinical routine involves dense baseline measurements, while subsequent follow-up assessments - particularly in rare diseases such as epidermolysis bullosa (EB) - are sparse and irregularly spaced. This distorted observation process prevents reliable inference of the underlying disease course over time, even if there is information about the relationship between the biomarkers. To overcome the information-sparse small-data regime we augment the limited measurements with a Neural Ordinary Differential Equation (Neural ODE), which learns a continuous-time latent representation capable of reconstructing distinct patient trajectories despite the sparse and irregular sampling.

Methods

Using an ODE based on prior expert knowledge as ground truth we generate realistic synthetic biomarker trajectories for several sub-groups with different initial conditions and dynamic parameters. To avoid bias toward a single dynamical model, synthetic data used to train the Neural ODE was sampled from prior distributions over ODE parameters, covering a realistic spectrum of inflammatory courses. A joint autoencoder-Neural ODE architecture is trained: the encoder maps each biomarker vector into a latent representation, while a latent-space Neural ODE regularizes the temporal evolution of these latent states.

Results

We demonstrate under which conditions of the observation process Neural ODEs can still reliably reconstruct latent patient trajectories and separate subgroups in the latent space. Our simulations examine how reconstruction error, subgroup separability, and trajectory stability change as measurement sparsity and irregularity increase. Furthermore, we identify critical thresholds for sampling density and missingness patterns beyond which identifiability of the underlying dynamics breaks down.

Discussion

Using synthetic trajectories generated from a realistic ODE based on prior expert knowledge as training data, our joint autoencoder-Neural ODE architecture learns nonlinear dynamics and flexible latent representations. The combination of realistic data and nonlinear continuous-time modeling offers advantages over other methods proposed for irregular and noisy time series like linear ODEs, RNNs like LSTM, and latent ODE variants when handling the severe sparsity typical of rare disease biomarker data.

[01-P2-15] Large Language Model-Assisted CT Protocol Selection: A Comparative Study of Fine-tuning vs Zero-Shot Reasoning

Presenter: Philipp Arnold

Authors: Philipp Arnold [1]; Sebastian Walter [2]; Elmar Kotter [1]; Hannah Bast [3]; Thomas Stein [1]

Affiliations: [1] Department of Radiology, Medical Center - University of Freiburg, Freiburg im Breisgau; [2] University of Freiburg, Freiburg im Breisgau; [3] Department of Computer Science, University of Freiburg, Germany, Freiburg im Breisgau

Introduction

Indication-specific computed tomography (CT) protocol selection is an essential radiology workflow step that requires balancing diagnostic performance against radiation exposure and the risks associated with iodinated contrast media. Protocol assignment depends on clinical indication, patient history, and imaging context, making the process complex and susceptible to inter-reader variability. Large language models (LLMs) may provide decision support by automatically recommending

CT protocols from free-text clinical information. This study evaluated the ability of LLMs to support CT protocol selection by comparing supervised fine-tuning with zero-shot reasoning approaches.

Methods

We retrospectively analyzed 20,000 CT examinations from a tertiary-care academic hospital. Each examination included the clinical indication, relevant patient history, and a standardized reference protocol containing body region and intravenous contrast media use. Fine-tuned Qwen-2.5 models with 3B, 7B, 14B, and 32B parameters were trained to predict the correct protocol label. In parallel, zero-shot Qwen-3 models (4B, 8B, 14B, and 32B) received structured prompts containing the same clinical information and protocol definitions during inference. Additional experiments evaluated the inclusion of prior imaging reports. Model performance was compared with protocol selections made independently by three board-certified radiologists. Evaluation metrics included exact-match accuracy, body-part accuracy, contrast-media accuracy, and top-3 accuracy.

Results

Fine-tuned models demonstrated the strongest overall performance, achieving exact-match accuracies of 69.4%-75.2% and top-3 accuracies of 84.4%-86.4%. The best-performing model, Qwen-2.5 14B, achieved 75.2% exact-match accuracy, 93.6% body-part accuracy, and 94.2% contrast-media accuracy. Zero-shot models achieved radiologist-level performance for individual decision components, with body-part accuracies of 91.0%-93.0% and intravenous contrast-media accuracies of 90.0%-95.5%. However, their exact-match accuracy remained lower at approximately 53.0%-55.5%. Incorporating prior imaging reports improved contrast-media decision accuracy in several zero-shot settings.

Discussion

LLMs demonstrated strong potential for assisting indication-specific CT protocol selection. Fine-tuning on institutional data achieved the highest overall protocol-selection accuracy and outperformed radiologists on exact protocol matching. In contrast, zero-shot reasoning provided competitive performance without task-specific training and showed flexibility for integrating prior imaging information. These findings suggest that both approaches may support radiology workflow optimization, although prospective multicenter validation is required before clinical implementation.

[01-P2-16] Agentic ODE Discovery and Parameter Distribution Inference from Summary Statistics for Rare Diseases

Presenter: Hanning Yang

Authors: Hanning Yang [1]; Meropi Karakioulaki [2]; Lennart Purucker [3, 4]; Tim Litwin [1]; Cristina Has [2]; Moritz Hess [1]

Affiliations: [1] Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center - University of Freiburg, Freiburg; [2] Department of Dermatology, Medical Faculty and Medical Center - University of Freiburg, Freiburg; [3] Prior Labs, Freiburg im Breisgau; [4] University of Freiburg, Freiburg, Germany, Freiburg

Introduction

Mechanistic modeling via ordinary differential equations (ODEs) offers interpretable descriptions of complex biological dynamics and is particularly valuable in rare disease research. However, rare diseases present a compounding set of challenges: model structure and parameters are typically unknown, individual-level data is scarce and noisy, and privacy constraints limit data sharing. These conditions demand methods that explicitly handle both structural uncertainty over which ODE governs the system, and parametric uncertainty over heterogeneous patient populations.

Methods

We present AgentODE, an end-to-end agentic framework that jointly discovers ODE structures and infers parameter distributions solely from population-level summary statistics. A large language model (LLM) proposes candidate ODE structures through an outer discovery loop, while a parameter inference agent iteratively refines parameter distributions via an inner diagnosis-update loop. At each iteration, current parameter distributions are used to simulate synthetic trajectories. The agent then updates parameter distributions based on discrepancies between summary statistics from synthetic and empirical trajectories. An experience buffer with dynamic memory guides subsequent structure proposals, enabling structured exploration under uncertainty.

Results

We evaluate AgentODE on benchmark problems and two clinical datasets. One includes only 231 observations across 46 patients from a recessive dystrophic epidermolysis bullosa (RDEB) cohort . AgentODE achieves performance comparable to baselines with full individual-level data access, while operating exclusively on summary statistics. Critically, on RDEB, AgentODE recovers a mechanistically consistent ODE structure aligned with an independently developed expert model, whereas baselines, despite lower prediction error, recover implausible or degenerate structures. This reveals that in sparse, noisy settings, predictive performance alone is an insufficient criterion for mechanistic model quality.

Discussion

AgentODE demonstrates that population-level summary statistics, rather than being merely a privacy-preserving compromise, provide a more robust signal for uncertainty-aware structure discovery by condensing population patterns while mitigating individual-level noise. This opens a principled pathway for mechanistic modeling in rare diseases where data scarcity, heterogeneity, and privacy constraints have traditionally precluded such analyses.

[01-P2-18] nanoTabPFN: A Lightweight and Educational Reimplementation of TabPFN

Presenter: Johannes Hog

Authors: Alexander Pfefferle [1]; Johannes Hog [2]; Lennart Purucker [3]; Frank Hutter [3]

Affiliations: [1] Institut für Informatik, Albert-Ludwigs-Universität Freiburg, Freiburg; [2] Institut für Informatik, Albert-Ludwigs-Universität Freiburg, Freibrug; [3] Institut für Informatik, Albert-Ludwigs-Universität Freiburg, Freiburg im Breisgau

Introduction

Recent advances in foundation models designed for structured data have significantly transformed the landscape of tabular machine learning. Despite their impressive capabilities, understanding what makes these models effective remains an open challenge. Current open-source implementations tend to be sprawling and opaque, often comprising tens of thousands of lines of code with little accompanying documentation. This complexity creates a steep learning curve and limits the ability of newcomers to experiment with or build upon these systems.

Methods

We present nanoTabPFN, a compact and readable reimplementation inspired by the TabPFN v2 architecture. Our approach prioritizes clarity and minimalism, offering a concise training pipeline that operates on pre-generated synthetic datasets. The design choices throughout emphasize transparency and ease of modification over exhaustive feature coverage.

Results

When evaluated in a limited data regime, our implementation reaches performance on par with established machine learning methods after roughly one minute of training on a single GPU. Compared to the original TabPFN v2 training procedure, this corresponds to a roughly five-order-of-magnitude reduction in compute time.

Discussion

These findings suggest that meaningful experimentation with tabular foundation models does not require access to large-scale infrastructure. By offering a lightweight and approachable codebase, nanoTabPFN lowers the barrier for students and researchers who wish to explore, teach, or iterate on this class of models without prohibitive resource constraints.

Poster Session: P3 - SmallData Uncertainty

[01-P3-01] Assessing the Impact of Model Assumptions in Network Meta-Regression: A Simulation Study (Na Kwarteng, G Schwarzer, A Nikolakopoulou, T Evrenoglou)

Presenter: Nana-adjoa Kwarteng

Authors: Nana-adjoa Kwarteng [1]; Guido Schwarzer [2]; Julia Müller [3]; Adriani Nikolakopoulou [4, 2]; Theodoros Evrenoglou [2]

Affiliations: [1] Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center-University of Freiburg, Freiburg im Breisgau; [2] Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center-University of Freiburg, Freiburg; [3] University of Freiburg, Faculty of Medicine, Freiburg; [4] Department of Hygiene, Social-Preventive Medicine and Medical Statistics, School of Medicine, Aristotle University of Thessaloniki, Thessaloniki

Introduction

Network meta-regression (NMR) models extend network meta-analysis (NMA) by simultaneously synthesizing evidence on multiple treatments while adjusting for effect modifiers. By accounting for effect modification, NMR can reduce between-study heterogeneity relative to NMA. However, available modeling assumptions in NMR lead to several models that address related but non-identical research questions, complicating model selection. The consequences of model misspecification under varying network structures, levels of heterogeneity, and interaction assumptions remain unclear.

Methods

To evaluate the performance of alternative NMR models and provide practical recommendations, we conducted a simulation study, varying key network characteristics, including network sparsity, study designs, and heterogeneity magnitude. NMR was examined under frequentist and Bayesian frameworks. We considered the standard NMA model without interactions, as well as four NMR parameterizations differing in across-comparison interaction assumptions (common vs. independent interactions) and interaction consistency assumptions (with or without consistency).

Results

Combining network characteristics with true model specifications resulted in 120 simulation scenarios. In fully connected networks of two-arm studies with low heterogeneity and without effect modification, all NMR models demonstrated low bias and mean squared error. In contrast, NMA models overestimated treatment effects and heterogeneity in the presence of effect modification. NMR models with independent across-comparison interactions generally achieved nominal coverage and low bias when the data-generating mechanism for interactions was consistent and common across comparisons. However, their coverage deteriorated in sparse networks with between-study heterogeneity. In corresponding scenarios, models assuming common across-comparison interactions were more robust to sparsity and heterogeneity but yielded conservative inference for interaction effects. In networks containing multi-arm studies, NMR models lacking interaction consistency showed sub-nominal coverage in their corresponding data-generating mechanisms.

Discussion

Our findings reinforce that ignoring treatment-covariate interactions by relying on standard NMA may yield biased treatment effect and heterogeneity estimates. In the presence of effect modification, NMR models are preferable and require careful consideration of modeling assumptions, anticipated patterns of effect modification, and the structure of the evidence network. Models assuming consistent interactions are advantageous in networks with multi-arm studies. Sparse networks often necessitate common-interaction assumptions. These findings provide guidance for planning and interpreting network meta-analyses and meta-regressions in complex treatment networks.

[01-P3-03] Before and beyond the mixing time: New approximations for additive functionals of stationary Gauss-Markov processes.

Presenter: Gabriele Bellerino

Authors: Gabriele Bellerino [1]; Angelika Rohde [1]; Peter Pfaffelhuber [1]; Aljaž Godec [2]

Affiliations: [1] Department of Mathematical Stochastics, University of Freiburg, Freiburg im Breisgau; [2] Institute of Physics, University of Freiburg, Freiburg im Breisgau

Introduction

Whereas classical invariance principles for ergodic Markov chains address the situation in which the time horizon of observations is much larger than the mixing time, the quality of approximation is questionable when this is not the case anymore - even when starting the Markov chain in the invariant law. In this project, this small-sample situation is thoroughly addressed for Gauß-Markov processes, especially in terms of quantitative distributional approximations.

Methods

The adopted methods are: Wiener chaos decomposition of the additive functional and spectral representation of the triangular array.

Results

To approach the small-sample effect in an asymptotic framework, we study the additive functional along a triangular array of stationary processes with diverging mixing times. We prove quantitative and functional limit theorems for additive functionals along triangular arrays of stationary Gaussian Markov processes when the mixing time scales sub-, super- and proportionately to the number of observations.

Discussion

Our major finding is a phase-transition when the number of observations scales proportionately to the mixing time, together with the identification and interrelation properties of the emerging new limit processes at and before the mixing time. This work forms the basis for ongoing follow-up work: Gabriele Bellerino, Angelika Rohde, and Peter Pfaffelhuber (C01) are currently extending the results to general reversible Markov processes. Moreover, the obtained quantitative limit theorems provide groundwork for the new Project 03 by Angelika Rohde and Aljaž Godec (New 03).

[01-P3-05] Improving Biogeographical Ancestry Prediction Through Machine Learning-Based Classification and Marker Selection

Presenter: Carola S. Heinzel

Authors: Carola S. Heinzel [1]; Lennart Purucker [2]; Julika Schmitz [3]; Jana Naue [4]; Theresa Maurer [1]; Zoe Lange [5]; Frank Hutter [6]; Sabine Lutz-Bonengel [3]; Holger Dette [5]; Peter Pfaffelhuber [1]

Affiliations: [1] Department of Mathematical Stochastics, University of Freiburg, Freiburg im Breisgau; [2] Machine Learning Lab, University of Freiburg, Freiburg im Breisgau; [3] Institute of Legal Medicine, Medical Center, University of Freiburg, Freiburg im Breisgau; [4] Institute of Forensic Medicine, Medical Center - University of Freiburg, Freiburg im Breisgau; [5] Ruhr-Universität Bochum, Bochum; [6] Institut für Informatik, Albert-Ludwigs-Universität Freiburg, Freiburg im Breisgau

Introduction

The classification of individuals into ancestral populations, i.e., inferring biogeographical ancestry (BGA) from DNA traces, is an important task in forensic genetics. Tools such as SNIPPER or the Admixture Model are widely used in forensic practice. However, they have not yet been systematically compared to classifiers widely used in other disciplines. Moreover, while established methods provide strong intercontinental classification, prediction accuracy within Europe remains limited. A major factor contributing to this limitation is the set of genetic markers used for classification. Recent advances in machine learning offer new opportunities to improve both ancestry classification and the selection of ancestry-informative markers (AIMs).

Methods

We investigated machine learning approaches for BGA prediction at continental and intra-continental levels using published genomic datasets. First, the tabular foundation model TabPFN was benchmarked against established forensic classifiers. Second, a machine learning-based marker selection pipeline using permutation importance and deep learning models was applied to identify optimized AIM panels for intra-European classification. Marker sets containing 50-200 markers were evaluated using repeated cross-validation and compared with the VISAGE Enhanced Tool and allele frequency-based approaches. Third, we developed a method that detects, if the ancestral population of the trace is represented in the training data set.

Results

TabPFN consistently achieved the best overall classification performance across accuracy, ROC AUC, and log loss. At the continental level, classification accuracy increased from 84% using SNIPPER to 93% using TabPFN. For intra-European classification, accuracy increased from 43% to 48% across ten populations. Optimized marker selection further improved intra-continental prediction. For four European populations, accuracy increased from 68.2% using 104 VISAGE-selected markers to 73.7% with 100 newly selected markers and to 82.3% with 200 markers.

Discussion

Our findings demonstrate that machine learning improves both classification performance and marker selection for forensic ancestry prediction. In particular, tailored marker panels markedly enhance intra-European classification. The combination of advanced classifiers and optimized ancestry-informative markers enables more precise differentiation between genetically similar populations and highlights the potential of machine learning for fine-scale forensic ancestry analysis across different geographic regions.

[01-P3-07] End-to-End Compression for Tabular Foundation Models

Presenter: Guri Zabergja

Authors: Guri Zabergja [1]; Rafiq Kamel [2]; Arlind Kadra [2]; Christian Frey [2]; Josif Grabocka [2]

Affiliations: [1] University of Freiburg, Freiburg; [2] University of Technology Nuremberg, Nuremberg

Introduction

The long-standing dominance of gradient-boosted decision trees for tabular data has recently been challenged by in-context learning tabular foundation models. In-context learning methods fit and predict in one forward pass without parameter updates by leveraging the training data as context for predicting on query test points. While recent tabular foundation models achieve state-of-the-art performance, their transformer architecture based on the attention mechanism has quadratic complexity regarding dataset size, which in turn increases the overhead on training and inference time, and limits the capacity of the models to handle large-scale datasets.

Methods

In this work, we propose TACO, an end-to-end tabular compression model that compresses the training dataset in a latent space.

Results

We test our method on the TabArena benchmark, where our proposed method is up to 94x faster in inference time, while consuming up to 97% less memory compared to the state-of-the-art tabular transformer architecture, all while retaining performance without significant degradation. Lastly, our method not only scales better with increased dataset sizes, but it also achieves better performance compared to other baselines.

Discussion

These results show that learning to compress the in-context training set into a compact latent representation effectively overcomes the quadratic scaling bottleneck of attention-based tabular foundation models, while the compression rate offers a principled knob for trading accuracy against latency and memory. Since TACO is modular and supports chunking to over a million rows, latent context compression is a promising design principle for making tabular foundation models practical at real-world scale.

[01-P3-09] Efficacy of psychotherapy, pharmacotherapy, vs. their combination in chronic depression: a systematic review and network meta-analysis using aggregated and individual patient data

Presenter: Julia Müller

Authors: Julia Müller [1]; Moritz Elsaesser [1]; Theodoros Evrenoglou [2]; Pim Cuijpers [3, 4]; Orestis Efthimiou [5]; Daniel N. Klein [6]; Martin B. Keller [7]; Toshi A. Furukawa [8]; Elisabeth Schramm [9]

Affiliations: [1] Department of Psychiatry and Psychotherapy, Medical Center - University of Freiburg, Faculty of Medicine, Freiburg; [2] Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center-University of Freiburg, Freiburg; [3] Department of Clinical, Neuro and Developmental Psychology, WHO Collaborating Centre for Research and Dissemination of Psychological Interventions, Vrije Universiteit Amsterdam, Amsterdam; [4] Babeș-Bolyai University, International Institute for Psychotherapy, Cluj-Napoca; [5] Institute of Social and Preventive Medicine, University of Bern, Bern; [6] Department of Psychology, Stony Brook University, Stony Brook, New York; [7] Department of Psychiatry and Human Behavior, The Warren Alpert Medical School of Brown University, Providence; [8] Department of Health Promotion and Human Behavior, Kyoto University Graduate School of Medicine/School of Public Health, Kyoto; [9] Department of Psychiatry and Psychotherapy, Medical Center - University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg

Introduction

Chronic depression represents a common, highly disabling disorder. Several randomized controlled trials investigated the effectiveness of psychological, pharmacological, and combined treatments for chronic depression. This is the first overarching systematic review and network meta-analysis based on aggregated and individual patient data (IPD-AD-NMA) comparing the efficacy and acceptability of various treatment options for chronic depression.

Methods

We searched the Cochrane Library, MEDLINE (via Ovid), PsycINFO, PsycArticles, Web of Science, and the Metapsy database to identify randomized controlled trials (RCTs) investigating treatments for adults with chronic depression. The primary outcome was clinician-rated depression severity at post-treatment. Two reviewers independently screened studies and assessed risk of bias using the Cochrane Risk of Bias 2 tool. Individual participant data (IPD) were requested from all eligible studies and combined with aggregate data (AD) when IPD were unavailable. Outcomes and treatment categories were harmonized across studies, and missing IPD were addressed using multilevel multiple imputation. Bayesian one-stage random-effects multilevel network meta-analysis (NMA) and network meta-regression (NMR) models were conducted in R to estimate comparative and individualized treatment effects while accounting for heterogeneity and inconsistency across the evidence network.

Results

We screened 3,971 studies for eligibility, selecting 225 for full-text review. 94 studies from 77 trials met inclusion criteria and 69 studies from 56 trials (IPD:42, AD: 14) were included in the analysis. Results from the analysis will be presented.

Discussion

This systematic review and meta-analysis aims to clarify and quantify the impact of various treatment options and the role of predictive and modifying factors in chronic depression, providing insights that can inform personalized treatment strategies for patients with chronic depression.

[01-P3-11] GUIDE-SAFE: Uncertainty-Aware Release Control for CKD/AKI Medication-Safety Decision Support

Presenter: Yaser Kord

Authors: Moritz Hess [1]

Affiliations: [1] Universität Freiburg, Freiburg

Introduction

Clinical AI systems for medication safety are usually evaluated as prediction models, but deployable decision support requires a stricter question: when is the available evidence safe enough to show to a human reviewer? This is a small-data uncertainty problem in CKD/AKI care, where renal function, potassium, medication timing, and AKI instability are often missing, stale, or clinically conflicting.

Methods

We developed GUIDE-SAFE, an offline release-controlled medication-safety decision-support prototype for CKD/AKI-related review cards. Each patient-medication context is assigned one of three actions: release, defer, or tension. Release denotes sufficiently consistent evidence for pharmacist review defer denotes missing, stale, or unstable evidence and tension denotes clinically relevant but conflicting evidence. The prototype combines deterministic medication-safety rules with a calibrated machine-learning controller and retrospective ICU data. The current validation phase, developed with medication-safety expertise, refines the card schema, action definitions, workflow placement, and blinded pharmacist adjudication rubric.

Results

The implemented prototype produces structured review cards containing medication exposure, renal-state evidence, potassium evidence, temporal uncertainty, AKI instability, missingness indicators, and reason codes. The planned validation evaluates unsafe release, useful release coverage, defer/tension burden, and failure modes such as stale creatinine, missing potassium, ambiguous medication timing, nephrotoxin stacking, and unstable AKI. The clinical workflow under evaluation treats the card as a review artifact, not an autonomous recommendation.

Discussion

GUIDE-SAFE reframes medication-safety AI as controlled release under uncertainty rather than direct prediction. The contribution is a practical assurance layer for small-data clinical AI: uncertain or conflicting cases are not forced into recommendations but routed to defer or tension states for human review. This design supports traceable evidence, calibrated abstention, pharmacist oversight, and deployment-readiness evaluation before any prospective clinical use.

[01-P3-13] Feature Selection with and for Tabular Foundation Models

Presenter: Dominika Matus

Authors: Dominika Matus [1]; Lennart Purucker [2]

Affiliations: [1] Institut für Informatik, Albert-Ludwigs-Universität Freiburg, Freiburg im Breisgau, Freiburg im Breisgau; [2] Institut für Informatik, Albert-Ludwigs-Universität Freiburg, Freiburg im Breisgau

Introduction

Tabular foundation models (TFMs) achieve strong performance on tabular prediction tasks but are typically restricted to a few hundred input features, whereas many real-world datasets, especially in the life sciences, contain thousands. Feature selection is a standard approach in such high-dimensional settings, yet its interaction with TFMs remains largely unexplored. This thesis investigates how feature selection affects the performance of TFMs on real high-dimensional tabular data and whether a TFM-based feature selector pretrained on synthetic priors can be competitive with well-established methods.

Methods

We curate seven high-dimensional real-world tabular datasets, predominantly from life-science domains, and evaluate a broad set of classical and deep feature selection methods, together with three TFMs (TabPFNv2, TabPFN-v2.5, and TabPFN-wide), as a step toward benchmarking TFMs on high-dimensional tabular data. In addition to these methods, we evaluate the TFMs without any feature selection to characterize their behavior on the full feature space. As a complementary contribution, we propose TabPFN-FS, a TFM-based feature selection method that attaches a small pretrained decoder to a frozen TabPFN-wide backbone.

Results

The results suggest that reducing high-dimensional tabular data with feature selection yields only modest improvements in downstream performance compared to applying TFMs directly to the full feature set. TabPFN-FS achieves performance on par with the best-performing method, Recursive Feature Elimination, while offering the computational advantages of a zero-shot setting.

Discussion

Performance differences between the three TFMs under the same feature selection methods are generally small, indicating that, for the datasets considered, these models already extract most of the available predictive signal from the high-dimensional feature space and that feature selection can at best provide incremental gains over this strong baseline.

STRATOS Symposium

Organizers

Anne-Laure Boulesteix, Georg Heinze, Willi Sauerbrei

Focus

Symposium Title: Application of Statistical Methods Needs to Improve - On the Critical Role of Guidance for Analysis and Knowledge Translation.

The international STRATOS (STRengthening Analytical Thinking for Observational Studies) initiative was founded in 2013 at the ISCB Annual Conference and organizes a mini-symposium at each annual meeting on methodological aspects of planning and analysing data in observational studies.

This time, STRATOS is organizing a mini-symposium for two different target groups on October 1st. The morning session (09:00-12:30, Methodological aspects) is designed in the usual format of a half-day meeting with several shorter talks on behalf of TGs and panels. In the afternoon (13:30-17:00, Practical aspects of medical research), STRATOS offers, for the first time, a workshop for clinicians and data analysts with less training and experience in statistical topics. Gary Collins (University of Birmingham, UK) and Matthias Briel (University Hospital Basel, Switzerland) will give keynote presentations.

Agenda

Session 1: Methodological Aspects, 09:00-10:30 and 11:00-12:30.

Session 1A.

09:00-09:10 Willi Sauerbrei: Introduction.

09:10-09:35 James R. Carpenter: P-values and hypothesis testing: beyond polemics to practical solutions.

09:35-10:00 Michal Abrahamowicz: How important are the Hazards of Hazard Ratios?

10:00-10:25 Anne C. M. Thiébaut: Methods for Adjusting for Covariate Measurement Error in Flexible Modelling of Functional Form: Results of a Blinded, Controlled Neutral Comparison Simulation Study.

Session 1B.

11:00-11:25 Marianne Huebner: Statistical analysis plan with initial data analysis (SAPI): Validating the SAPI checklist in analysis projects.

11:25-11:50 Doranne Thomassen: Comparing five frameworks that may be used to define estimands: friends not foes.

11:50-12:15 Els Goetghebeur: Sensitivity analyses for missing data in observational studies: A practical guide for planning, conducting and reporting under MAR and beyond.

12:15-12:30 Discussion 1.

Session 2: Practical Aspects of Medical Research, 13:30-15:00 and 15:30-17:00.

Session 2A.

13:30-13:40 Willi Sauerbrei: Introduction.

13:40-14:10 Gary Collins: The importance of methodology and transparency in predictive AI.

14:10-14:27 Riccardo De Bin: From traditional statistical modelling to machine learning algorithms: a perspective from the STRATOS TG9 high-dimensional data viewpoint.

14:27-14:44 Ben van Calster: Performance evaluation of predictive AI models to support medical decisions: overview and guidance.

14:44-15:01 Peggy Sekula: Designing Prognostic Factor Studies: Key Concepts and Practical Guidance.

Session 2B.

15:30-16:00 Matthias Briel: Meta-research to improve evidence generation and synthesis.

16:00-16:15 Theresa Ullmann: The Problem with Univariable Selection in Regression Modelling - and What to Do Instead.

16:15-16:30 Saskia le Cessie: Guidelines to the design, analysis and interpretation of patient-reported outcomes in cancer clinical trials.

16:30-16:45 Willi Sauerbrei: An overview and categorization of papers published in statistical series of medical journals.

16:45-17:00 Discussion 2.

STRATOS Symposium Session 1A: Methodological Aspects

Thu, 01.10.26, 09:00–10:30, Room: Runder Saal

Chair: Willi Sauerbrei

Session 1A.

09:00-09:10 Willi Sauerbrei: Introduction.

09:10-09:35 James R. Carpenter: P-values and hypothesis testing: beyond polemics to practical solutions.

09:35-10:00 Michal Abrahamowicz: How important are the Hazards of Hazard Ratios?

10:00-10:25 Anne C. M. Thiébaut: Methods for Adjusting for Covariate Measurement Error in Flexible Modelling of Functional Form: Results of a Blinded, Controlled Neutral Comparison Simulation Study.

P-values and hypothesis testing: beyond polemics to practical solutions

09:10-09:35

Presenter/Author: James Carpenter

Affiliations: London School of Hygiene & Tropical Medicine, London

Despite recent articles highlighting concerns about the ubiquitous use and misleading interpretation of p-values, a glance at the observational research literature confirms problems persist. This prompted the STRATOS initiative to explore practical proposals to improve practice. Building on recent literature, we distinguish three broad goals for observational data: building descriptive models (e.g. to describe associations and develop hypotheses), building prediction models and addressing causal questions. For each goal, we discuss study planning considerations, and the key tasks p-values can be used for (e.g. evaluating hypotheses as tuning parameter in model selection) using examples to highlight key considerations for good practice. These include study registration, thorough initial data analysis, an increased emphasis on reporting of the methodological steps, as well as structured reporting of the results. We argue all are necessary for reproducible research.

How important are the Hazards of Hazard Ratios?

09:35-10:00

Presenter: Michal Abrahamowicz

Authors: Michal Abrahamowicz [1]; Marie-Eve Beauchamp [2]; Emily Roberts [3]; Jeremy Taylor [4]

Affiliations: [1] Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal; [2] McGill University, Montral; [3] University of Iowa, USA, Iowa; [4] University of Michigan, Michigan

In time-to-event data analyses, the hazard ratio (HR) is a frequently used metric of the strength of the association. In a highly cited commentary, M. Hernán criticized using of Cox proportional hazards (PH) model-based HRs by emphasizing that they suffer from the 'built-in selection bias [Epidemiology 2010]. To illustrate this bias, he presented the results of a single real-world randomized trial and argued that the estimated decreases in HR with increasing follow-up time are likely simply due to a latent frailty and, thus, should not be interpreted as the evidence of the true (causal) treatment effect diminishing over time. This article led to major concerns about the validity of Cox PH regression analyses. We rely on comprehensive simulations and real-world analyses to revisit this controversial issue First, we demonstrate that the built-in selection bias is usually quite modest, unless the unmeasured frailty has a very strong impact on survival. Then, in dedicated simulations that closely mimic the data structure of the trial discussed by Hernán, we demonstrate that the observed time-dependent changes in treatment HR were practically impossible to reflect just an unmeasured risk factor, even with an extremely strong effect. This finding calls for alternative explanations, that likely include a combination of decreasing treatment adherence and biological changes associated with prolonged treatment. Overall, our simulation results provide strong verifiable evidence that the concerns about the built-in selection bias, and about modeling and interpretation of HRs, are largely overstated.

Methods for Adjusting for Covariate Measurement Error in Flexible Modelling of Functional Form: Results of a Blinded, Controlled Neutral Comparison Simulation Study

10:00-10:25

Presenter: Anne Thiebaut

Authors: Aris Perperoglou [1]; Mohammed Sedki [2]; Anne Thiebaut [3]; Steve Ferreira Guerra [4]; Paul Gustafson [5]; Frank E. Harrel Jr. [6]; Willi Sauerbrei [7]; Michal Abrahamowicz [8]; Laurence Freedman [9]

Affiliations: [1] GSK, London; [2] Université Paris-Saclay, UVSQ, Inserm, Gustave Roussy, Villejuif; [3] Université Paris-Saclay, UVSQ, Inserm, Villejuif, France, Villejuif; [4] Centre de Recherche du Centre Hospitalier de l'Université de Montréal (CRCHUM), Montréal; [5] Department of Statistics, University of British Columbia, Vancouver; [6] Department of Biostatistics, Vanderbilt University School of Medicine, Nashville; [7] Institut für Medizinische Biometrie und Informatik, Universitätsklinikum Freiburg, Freiburg; [8] Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal; [9] Information Management Services Inc., Calverton, Maryland, USA, Calverton

Covariates in medical research are often measured with error, biasing estimates of exposure-outcome relationships, especially when these relationships are non-linear. This study compares methods for

measurement error correction combined with flexible regression modelling techniques in such non-linear settings. This blinded, controlled neutral comparison, multi-stage simulation project, a collaboration within the STRATOS initiative (Topic Groups 2 and 4), involved a Data Generation and Evaluation team and three Methods teams. These teams applied Bayesian methods, Multiple Imputation/Regression Calibration (MI/RC), and Simulation Extrapolation (SIMEX), combined with flexible modelling techniques (B-splines (BS), P-splines (PS), Fractional Polynomials (FP), and Natural Splines (NS)). Datasets featured a binary outcome, a continuous covariate with classical error, and a replicate substudy. The true non-linear functional form, covariate distribution, error variance, and error distribution were initially withheld. Stage 1 used 5 pilot datasets. Stage 2 expanded to 150 unique datasets by varying sample sizes, measurement error (ME) variance, error distribution (Normal, shifted-Gamma), and true functional forms. Stage 3 consisted of simulating independent replications of each combination to quantify the sampling variance of the estimates. Performance was assessed by log Mean Squared Error over pre-selected values of the covariate distribution. SIMEX methods consistently demonstrated the highest accuracy. P-splines, FPs, and NS generally outperformed BS, especially with SIMEX or Bayesian approaches. Following SIMEX, Bayesian methods (excluding BS) performed best, then RC (excluding BS), and MI. Bayesian BS combinations typically performed poorest, particularly with smaller samples. Accuracy generally improved with larger sample sizes and smaller ME. Linear relationships were estimated most accurately J-shaped forms were most challenging. Notably, SIMEX was less sensitive to increased ME magnitude and, unlike MI and Bayes, showed no substantial accuracy improvement with larger replication substudy sizes. This joint work emphasizes the relevance of neutral comparison studies to fairly evaluate statistical methods aimed at addressing a complex analytical challenge, and demonstrates their feasibility through a large collaborative project.

STRATOS Symposium Session 1B: Methodological Aspects

Thu, 01.10.26, 11:00–12:30, Room: Runder Saal

Chair: Georg Heinze

Session 1B.

11:00-11:25 Marianne Huebner: Statistical analysis plan with initial data analysis (SAPI): Validating the SAPI checklist in analysis projects.

11:25-11:50 Doranne Thomassen: Comparing five frameworks that may be used to define estimands: friends not foes.

11:50-12:15 Els Goetghebeur: Sensitivity analyses for missing data in observational studies: A practical guide for planning, conducting and reporting under MAR and beyond.

12:15-12:30 Discussion 1.

Statistical analysis plan with initial data analysis (SAPI): Validating the SAPI checklist in analysis projects

11:00-11:25

Presenter: Marianne Huebner

Authors: Carsten Oliver Schmidt [1]; Lara Lusa [2]; Georg Heinze [3]

Affiliations: [1] Universität Greifswald, Greifswald; [2] Faculty of Medicine, Institute of Biostatistics and Medical Informatics, University of Ljubljana, Ljubljana; [3] Medical University of Vienna, Wien

The SAPI statement (Statistical Analysis Plan for observational studies with Initial data analysis) provides recommendations for developing analysis plans for observational studies. It promotes a systematic process to improve completeness and transparency, and ultimately to better interpretability of research results for evidence-based health decision making. It can be used for descriptive, predictive, or causal research objectives. The SAPI checklist consists of 27 items comprising administrative information, project background, design and data, variables, main data analysis (MDA), initial data analysis (IDA), evaluation and updates, and supplement with plans for disseminating and archiving research outputs. We demonstrate these items in several examples of analysis plans and their updates following IDA including investigating functional forms in regression models or estimating learning curves. In each of these examples, updates of the SAP were implemented after IDA. These examples serve to demonstrate that a full pre-specification of an SAP is rarely possible without data examinations. The SAPI checklist and IDA framework guarantee that these looks are performed while strictly ignoring outcome-predictor associations and that their impact on the final analysis specification is transparently reported.

Comparing five frameworks that may be used to define estimands: friends not foes

11:25-11:50

Presenter: Doranne Thomassen

Authors: Nicholas Bakewell [1]; Doranne Thomassen [2]; Jonathan Bartlett [3]; Suzanne Cadarette [4]; Saskia Le Cessie [2]; Nan van Geloven [5]

Affiliations: [1] University of Toronto, Canada, Toronto; [2] Leiden University Medical Center, Leiden; [3] London School of Hygiene & Tropical Medicine, London; [4] University of Toronto, Toronto; [5] Leiden University Medical Center, the Netherlands, Leiden

There is a surge of interest in defining estimands, which translate a research question into a precise target quantity for estimation. The methodological literature is however divided over how best to define estimands, which may confuse researchers. To provide practical guidance, we offer a comparative overview of five key frameworks that may be used to define estimands: PICOT (Population, Intervention, Comparator, Outcome, Time), the ICH E9(R1) addendum on estimands, the Target Trial Emulation Framework, the Causal Roadmap, and the STRATOS (STRengthening Analytical Thinking for Observational Studies) Causal Inference Topic Group framework. After outlining each framework,

we demonstrate their similarities and divergence by applying them to a common research question. We then highlight key strengths and limitations of each framework for defining estimands. We find that each framework has unique strengths and limitations, reflecting its original purpose and scope. We advocate for an approach in which researchers thoughtfully choose and sometimes combine elements of frameworks to leverage their strengths. Using estimands in this way promotes greater clarity, minimizes ambiguity, and enhances the transparency and interpretability of study results. This presents joint work of members from STRATOS Topic Groups 7 (Causal Inference), 5 (Study design), and 4 (Measurement error and misclassification): Kelly Van Lancker (Ghent University, Belgium), Els Goetghebeur (Ghent University, Belgium), Pamela Shaw (University of Washington, USA), Emmanuelle Boutmy (Merck HealthCare KGaA, Germany), Tim Morris (Novartis Pharma UK Ltd, UK), Rima Izem (Novartis Pharma AG, Switzerland), Susan Halabi (Duke University, USA), Daniala L Weir (Utrecht University, the Netherlands), Marc Vandemeulebroecke (Bayer BCC AG, Switzerland)

Sensitivity analyses for missing data in observational studies: A practical guide for planning, conducting and reporting under MAR and beyond

11:50-12:15

Presenter: Els Goetghebeur

Authors: Dries Reynders [1]; Rheanna Mainzer [2]; Saskia Le Cessie [3]; James Carpenter [4]; Els Goetghebeur [5]; Katherine Lee [6]

Affiliations: [1] Ghent University, Ghent; [2] Murdoch Children's Research Institute, Melbourne, Australia, Melbourne; [3] Leiden University Medical Center, Leiden; [4] London School of Hygiene & Tropical Medicine, London; [5] Ghent University - Department of Mathematics, Computer Science and Statistics, Ghent; [6] Murdoch Children's Research Institute, Melbourne

In observational studies as in randomized trials with longitudinal outcomes, missing data are hard to avoid. Clear assumptions on the missing data mechanisms are then needed to allow for unbiased estimation of most target estimands. As these assumptions are typically untestable, sensitivity analyses come highly recommended. Allowing for missingness not at random (MNAR) can be complicated and time-consuming. It may furthermore arrive at bounds on the estimated estimand that are unduly wide. How to usefully accomplish a sensitivity analysis depends on the study design and target estimand. Little guidance is available on the planning, conduct and reporting of such sensitivity analyses, even in standard settings. In this talk we report on such guidance developed by TG1 to help ensure robust study conclusions, with a check list to aid in reporting these types of sensitivity analyses. We next list broader issues considered in collaboration between the missing data and causal inference topic groups. In cancer trials, missingness may coincide with intercurrent events and hit a non-positivity boundary. We discuss how recording reasons of missingness as a design feature can support more directed sensitivity analyses. We illustrate our findings and solutions, implementing several strategies in an observational study and a clinical trial with health-related quality of life outcomes. By thus supporting the uptake of sensitivity analyses we seek to contribute to improved research quality, transparency and reproducibility.

STRATOS Symposium Session 2A: Practical Aspects of Medical Research

Thu, 01.10.26, 13:30–15:00, Room: Runder Saal

Chair: Willi Sauerbrei

Session 2A.

13:30-13:40 Willi Sauerbrei: Introduction.

13:40-14:10 Gary Collins: The importance of methodology and transparency in predictive AI.

14:10-14:27 Riccardo De Bin: From traditional statistical modelling to machine learning algorithms: a perspective from the STRATOS TG9 high-dimensional data viewpoint.

14:27-14:44 Ben van Calster: Performance evaluation of predictive AI models to support medical decisions: overview and guidance.

14:44-15:01 Peggy Sekula: Designing Prognostic Factor Studies: Key Concepts and Practical Guidance.

The importance of methodology and transparency in predictive AI

13:40-14:10

Presenter: Willi Sauerbrei

Authors: Gary Collins

Affiliations: Department of Health Sciences, School of Health Sciences, college of Medicine and Health, University of Birmingham, Birmingham, United Kingdom, Birmingham

The transformative potential of predictive AI in healthcare hinges on building trust, which is predicated on transparency across the entire prediction model lifecycle. This talk will address the critical importance of transparency for fostering trustworthy AI in healthcare, from the initial stages of model development through rigorous evaluation in clinical trials. I will explore how predictive AI studies can obscure biases, hinder clinical adoption, and undermine patient safety. To combat these challenges, I will highlight some best practices for transparent AI development and evaluation, including the implementation of key reporting guidelines such as TRIPOD+AI. Furthermore, I will emphasize how they play a crucial role in reproducibility, critical appraisal, and open science in ensuring the robustness and reliability of AI-driven healthcare solutions. I will discuss the importance of openly sharing code, data, and model parameters, and enabling independent validation. By advocating for these principles, researchers can pave the way for a future where AI in healthcare is not only innovative but also ethically sound and demonstrably trustworthy.

From traditional statistical modelling to machine learning algorithms: a perspective from the STRATOS TG9 high-dimensional data viewpoint

14:10-14:27

Presenter: Riccardo De Bin

Authors: Riccardo De Bin [1]; Federico Ambrogi [2]; Lara Lusa [3]; Jörg Rahnenführer [4]; Willi Sauerbrei [5]

Affiliations: [1] University of Oslo, Oslo; [2] Università degli Studi di Milano, Milan; [3] Faculty of Medicine, Institute of Biostatistics and Medical Informatics, University of Ljubljana, Ljubljana; [4] Department of Statistics, TU Dortmund University, Dortmund; [5] Institut für Medizinische Biometrie und Informatik, Universitätsklinikum Freiburg, Freiburg

The rise of machine learning, driven by technological and algorithmic advances, has brought with it the promise of data-driven solutions to a wide range of problems, particularly when high-dimensional data from multiple sources are available. For example, in the field of precision medicine, information from various omics data can be integrated to identify specific disease subtypes that may benefit from targeted treatments. The traditional statistical approach to data analysis has been contested by the problem-oriented attitude of machine learning. Nowadays, the classical view of the two cultures, although still fascinating, needs to be replaced by a new vision in which inference and algorithms coexist to address modern challenges: Statistical modelling and machine learning represent a continuum that may help analysts in different contexts or in specific problems. In this talk, we provide a perspective on this issue from the STRATOS TG9 High-dimensional data viewpoint, focusing on similarities, differences, and common requirements (e.g., reproducibility, reporting) between the two approaches.

Performance evaluation of predictive AI models to support medical decisions: overview and guidance

14:27-14:44

Presenter: Ben Van Calster

Authors: Ben Van Calster [1]; Ewout Steyerberg [2]

Affiliations: [1] Department of Development and Regeneration, KU Leuven, Leuven; [2] University Medical Center Utrecht, Utrecht University, Utrecht

Selecting appropriate performance measures is essential for clinical AI predictive models intended to support medical decisions. The use of suboptimal performance measures may lead to implementing AI models that suggest detrimental clinical decisions. We illustrate this challenge using the ADNEX model, which predicts the probability of malignancy in women with an ovarian tumor. We externally

validate ADNEX (n=894) and a recalibrated version tailored to the population at hand. Although recalibration improved the model, some measures suggested worse performance (e.g. F1). We evaluate the merits of classic and contemporary performance measures and plots for models with a binary outcome. We consider five performance domains (discrimination, calibration, overall, classification, utility) and identify two key characteristics for selecting a performance measure: 'properness' and 'focus'. Only 17 of 32 assessed measures possessed both characteristics, while one possessed neither (F1). Several problematic measures are becoming increasingly popular, making it challenging to improve model validation practices. To provide guidance, we recommend the following measures and plots as essential to report: c-statistic (or AUROC), calibration plot, a clinical utility measure with a decision curve, and a plot showing the distribution of estimated risks. Published as: Van Calster B, Collins GS, Vickers AJ, Wynants L, Kerr KF, Barreñada L, Varoquaux G, Singh K, Moons KGM, Hernandez-Boussard T, Timmerman D, McLernon DJ, van Smeden M, Steyerberg EW, for TG6 of the STRATOS initiative. Performance evaluation of predictive AI models to support medical decisions: overview and guidance. Lancet Digital Health 20257:100916.

Designing Prognostic Factor Studies: Key Concepts and Practical Guidance

14:44-15:00

Presenter: Peggy Sekula

Authors: Peggy Sekula [1]; Suzanne Cadarette [2]; Mitchell H Gail [3]

Affiliations: [1] Institute of Epidemiology and Prevention, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg im Breisgau; [2] University of Toronto, Toronto; [3] Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland, USA, Rockville

The investigation of factors that predict clinically relevant outcomes in diseased persons (prognostic factor research) is important to support the development of prognostic tools that can ultimately be used to tailor treatment decisions. As the output and quality of such studies are in need of improvement, the topic group 'TG5: study design' of the STRATOS initiative has developed a guidance document for general readership, intended to assist the design of studies on prognostic factors. The document also includes a glossary of terms and a list of general aspects to consider when designing a prognostic factor study. In this presentation, we will provide an overview of the aspects covered in the document and illustrate them using a number of examples.

STRATOS Symposium Session 2B: Practical Aspects of Medical Research

Thu, 01.10.26, 15:30–17:00, Room: Runder Saal

Chair: Anne-Laure Boulesteix

Session 2B.

15:30-16:00 Matthias Briel: Meta-research to improve evidence generation and synthesis.

16:00-16:15 Theresa Ullmann: The Problem with Univariable Selection in Regression Modelling - and What to Do Instead.

16:15-16:30 Saskia le Cessie: Guidelines to the design, analysis and interpretation of patient-reported outcomes in cancer clinical trials.

16:30-16:45 Willi Sauerbrei: An overview and categorization of papers published in statistical series of medical journals.

16:45-17:00 Discussion 2.

Meta-research to improve evidence generation and synthesis

15:30-16:00

Presenter/Author: Matthias Briel

Affiliations: University of Basel, Basel

Randomized clinical trials (RCTs) are critical for the generation of trustworthy evidence on clinical interventions, and individual participant data meta-analysis (IPDMA) is deemed the 'gold standard' for synthesizing randomized evidence to best inform clinical practice and policy. In this talk, I will present meta-research showing that one out of four RCTs is prematurely discontinued, mostly due to poor recruitment of participants and particularly affecting investigator-initiated RCTs. Results of one third of started investigator-initiated RCTs are not published in peer-reviewed publications or trial registries. Results of prematurely discontinued RCTs even remain unpublished in about 50% of cases. Non-publication of randomized evidence may seriously compromise evidence synthesis for clinical decision making alongside other issues such as low reporting quality of trials, heterogeneity of trial outcomes and subgroup definitions, or flawed trial data. While IPDMA is able to overcome many of these issues, it has its own challenges. I will present current meta-research on IPDMAs and discuss how the field of IPDMA has changed in the last two decades including the introduction of data sharing policies and the assessment of IPD for integrity issues. Finally, I will outline how prospective collaborative IPDMA can address the need for timely generation of high-quality evidence summaries.

The Problem with Univariable Selection in Regression Modelling - and What to Do Instead

16:00-16:15

Presenter: Theresa Ullmann

Authors: Theresa Ullmann [1]; Georg Heinze [2]; Franziska Kappenberg [3]; Marc Henrion [4]; Daniela Dunkler [1]

Affiliations: [1] Medical University of Vienna, Vienna; [2] Medical University of Vienna, Wien; [3] TU Dortmund, Bonn; [4] Malawi Liverpool Wellcome Research Programme, Blantyre, Malawi, Blantyre

Univariable selection, the screening of variables individually for statistical significance before inclusion in multivariable regression models, is one of the most widely used variable selection strategies in medical research. However, this approach lacks a theoretic justification, since a variable's association with an outcome in isolation does not reflect its role in a multivariable model. As a result, univariable selection can exclude important variables, retain spurious ones, and lead to biased estimates, invalid confidence intervals, and poor predictive performance. Three key mechanisms explain this failure. Masking occurs when correlated variables obscure each other's true effects, causing important variables to appear non-significant in univariable analyses. Amplification occurs when a variable appears significant in a univariable model only because it is correlated with another variable that has a true effect. Finally, even when variables are uncorrelated, univariable analyses may still fail to detect important variables due to a loss of precision: including multiple relevant variables in a multivariable model reduces the residual error and can reveal predictors that univariable analyses miss. We illustrate these problems using simulations and two real-world medical datasets (a breast cancer biomarker study and a study of lung health in schoolchildren). Combining univariable screening with subsequent backward elimination does not remedy the problem. We discuss alternatives to univariable selection and offer concrete recommendations for variable selection tailored to descriptive, predictive, and causal modelling goals.

Guidelines to the design, analysis and interpretation of patient-reported outcomes in cancer clinical trials

16:15-16:30

Presenter: Saskia Le Cessie

Authors: Saskia Le Cessie [1]; Doranne Thomassen [1]; Ahu Alanya [2]; Els Goetghebeur [3]

Affiliations: [1] Leiden University Medical Center, Leiden; [2] EORTC, Brussels, Belgium, Brussels; [3] Ghent University - Department of Mathematics, Computer Science and Statistics, Ghent

Since 2019, STRATOS is involved in the international SISAQOL-IMI consortium. This consortium aims to establish international standards in the analysis of patient reported outcomes (PRO) and health-related quality of life data in cancer clinical trials. Last year the SISAQOL-IMI consortium delivered its final recommendations for the design, analysis, interpretation and reporting of PROs in Randomised Clinical Trials (RCT) and Single Arm Studies (SAT). These recommendations can be found on https://www.sisaqol.org/. STRATOS contributed primarily in the work on SAT. In this talk we will give an overview of the results of this work, discussing questions and challenges in the design and analysis of single arm PRO studies.

An overview and categorization of papers published in statistical series of medical journals

16:30-16:45

Presenter: Willi Sauerbrei

Authors: Willi Sauerbrei [1]; Gary Collins [2]

Affiliations: [1] Institut für Medizinische Biometrie und Informatik, Universitätsklinikum Freiburg, Freiburg; [2] Department of Health Sciences, School of Health Sciences, college of Medicine and Health, University of Birmingham, Birmingham, United Kingdom, Birmingham

STRATOS compiled a collection of statistical articles published in series of medical journals between 1994 and 2025. These articles are primarily intended to be accessible to clinicians and researchers who have limited familiarity with statistical methodology. Applying a set of inclusion and exclusion criteria, we identified 35 yielding with 1188 articles for consideration. Statistical topics were classified into 36 categories (e.g., bias/systematic errors, confounding, sample size), seven of them with additional subcategories (e.g., Statistical modelling: a: model building, b: sensitivity analyses/check of model assumptions, c: other topics). To ensure consistency, 12 pairs of reviewers double-checked each paper. Members from all STRATOS Topic Groups, along with contributors from some panels assigned each article to up to three categories or recommended its exclusion. In total, 914 articles from 34 journals were categorized. The primary output of this project is a spreadsheet documenting, for each article, the number of pages, tables, figures, references and the assigned categories. It enables straightforward searching by topic and provides direct hyperlink access to articles of interest. This classification resource will support researchers in efficiently identifying methodological literature on a given statistical topic, drawing exclusively from statistical series published in medical journals.

Early Career Researcher Day

Organizers

ISCB Early Career in Biostatistics (ECB) Subcommittee and GMDS Early Career Representatives

Focus

Join the joint ISCB GMDS Early Career Research Day on October 1st, 2026. A full-day event for early career researchers in biostatistics, bioinformatics, epidemiology, and medical informatics. The program brings together the ISCB and GMDS early career communities for a day of talks, interactive sessions, and networking. Highlights include a talk on what AI means for the future of biostatistics, a talk on hallucination detection in large language models, speed networking, the Dilemma Game, a hands-on session, and a panel discussion. Whether you are a PhD student, postdoc, or early-stage professional, this day is designed to connect you with peers and mentors across both societies.

Agenda

Session 1

09:00-09:15 Opening remarks

09:15-10:00 Kristy Robledo: What Does AI Mean for the Future of Biostatistics

10:00-10:30 Speed Networking

10:30-11:00 Coffee Break

Session 2

11:00-11:30 Jasko Miebs: Hallucination Detection in Large Language Models

11:30-12:00 Ulrich Mansmann: Trustworthy Biostatistics in the Age of AI: Research Integrity as Professional Practice

12:00-12:30 Dilemma Game

12:30-13:00 Lunch Break

Session 3

13:30-14:30 Interactive Session

14:30-15:00 Panel Discussion - Voices from Early Career Researchers

Session Descriptions

Speed Networking: Attendees are divided into groups of about ten. Each participant has two to three minutes to give an elevator pitch or three-minute thesis style introduction of themselves and their work or research. A great opportunity to showcase your work while making new connections.

Dilemma Game: Building on the success of last year's session, this year's ECR Day includes an interactive session where all participants discuss the challenges statisticians face in conducting biostatistical research. We use the Dilemma Game app, developed by Erasmus University Rotterdam (EUR),

to pose questions and scenarios related to professionalism and integrity in research for discussion. We recommend that all attendees download the app before the mini-symposium.

Panel Discussion - Voices from Early Career Researchers: A key aim of the ECR Day is to provide a forum for discussing the challenges biostatisticians face. Rather than the traditional format where established researchers share their experiences with a largely listening audience, this session gives early-career researchers an opportunity to discuss the opportunities, challenges, and realities of building a career in statistics and biostatistics today. Topics may include research and collaboration, publishing, career development, balancing competing demands, and navigating transitions between academia, industry, and healthcare. The session aims to encourage open discussion, shared experiences, and peer learning, reflecting the collaborative spirit of ECR Day.

We invite all delegates registered for ECR Day who identify as early career researchers to put themselves forward to join the panel. A range of perspectives and backgrounds is welcome, and no previous panel experience is required. Please register your interest here: https://forms.office.com/e/L3csf-pyDHL

Early Career Researcher Day Session 1

Thu, 01.10.26, 09:00–10:30, Room: K5

Chair: Autumn Johnson, Judith Wodke

Session 1

09:00-09:15 Opening remarks

09:15-10:00 Kristy Robledo: What Does AI Mean for the Future of Biostatistics

10:00-10:30 Speed Networking

What Does AI Mean for the Future of Biostatistics?

09:15-10:00

Presenter/Author: Kristy Robledo

Affiliations: NHMRC Clinical Trials Centre, University of Sydney, Camperdown

Introduction: Generative artificial intelligence (AI) is increasingly influencing the practice of clinical trials biostatistics. These tools, which generate text, code, and other outputs through probabilistic prediction of language patterns, offer opportunities to enhance efficiency across routine tasks. However, their integration raises important considerations regarding validity, ethics, intellectual property, and the evolving role of statisticians. Methods: This work draws on a practical overview of generative AI applications in day-to-day biostatistical workflows, informed by current institutional practices and training initiatives, performed at the Australian Clinical Trials Alliance Statistics in Clinical Trials annual meeting in November 2025. Use cases were identified across common activities in clinical trials, including programming, documentation, communication, and scientific writing. Consideration was also given to differences between public and institutionally managed AI systems, and to emerging guidance on responsible use. Results: Generative AI can support a range of routine tasks, including drafting meeting summaries, generating and explaining statistical code, and assisting with manuscript development. These tools may improve efficiency by automating repetitive processes and aiding understanding of unfamiliar code. However, outputs typically rely on widely used methods and require careful validation, as errors in statistical methods, assumptions, or interpretation may occur. Institutional AI systems offer advantages for data security and intellectual property protection compared to public tools. Broader concerns include the inappropriate use of AI in peer review, confidentiality risks in clinical trials, and the need for transparency in reporting AI-assisted work. Discussion: Generative AI should be viewed as an assistive tool that complements, rather than replaces, statistical expertise. Its effective use requires critical evaluation, clear communication with collaborators, and adherence to ethical and institutional guidelines. While AI has the potential to reshape expectations of productivity and the skillset of biostatisticians, core competencies in problem-solving and statistical reasoning remain essential. As the field evolves, statisticians must engage proactively with AI technologies to ensure their responsible and appropriate application.

Early Career Researcher Day Session 2

Thu, 01.10.26, 11:00–12:30, Room: K5

Chair: Autumn Johnson, Max Behrens, Judith Wodke

Session 2

11:00-11:30 Jasko Miebs: Hallucination Detection in Large Language Models

11:30-12:00 Ulrich Mansmann: Trustworthy Biostatistics in the Age of AI: Research Integrity as Professional Practice

12:00-12:30 Dilemma Game

Hallucination Detection in Large Language Models

11:00-11:30

Presenter/Author: Jasko Miebs

Affiliations: University of Rostock, Rostock

Generative Large Language Models (LLMs) are a type of AI model that are becoming increasingly popular in various applications. These models are designed to generate a text response for any given input prompt. LLMs are trained on a large amount of data, before then being used for text generation. Gaps or contradictory information in the training data, as well as the model archi- tecture can lead to hallucinations. Hallucinations refer to the phenomenon where seemingly correct model outputs include false or inconsistent information. This poses a significant challenge in real-world applications, especially high-stakes ones like medical applications, where reliability is critical. We make use of Uncertainty Quantification (UQ) methods, to understand when these hallucinations occur. Uncertainty Quantification aims to develop methods and techniques to measure different types of uncertainty that may arise in machine learning models. We apply these methods to LLMs to identify when they are likely to produce hallucinations. Mathematical foundations of UQ methods deliver the reliability necessary in this field, and the results can be used to improve the perfor- mance of LLMs and make them more reliable in real-world applications. Many UQ methods can be easily implemented and are applicable to a wide range of models, without requiring a significant amount of additional computational resources. This makes them a promising approach for improving the reliability of LLMs and reducing the occurrence of hallucinations.

Trustworthy Biostatistics in the Age of AI: Research Integrity as Professional Practice

11:30-12:30

Presenter/Author: Ulrich Mansmann

Affiliations: Universität München, München

Research integrity is often discussed in relation to misconduct, regulation, or publication ethics. For biostatisticians, however, it is also a core element of daily professional practice. This talk introduces early-stage biostatisticians to research integrity as the craft of making statistical work transparent, reproducible, defensible, and useful for clinical and scientific decision-making. The presentation will follow the evidentiary chain of biostatistical work: clarifying the research question, defining estimands, writing statistical analysis plans, documenting deviations, producing reproducible code, supporting transparent reporting, and contributing to responsible data and code sharing. Particular attention will be given to situations in which biostatisticians face integrity-relevant decisions, such as handling missing data, choosing between competing analyses, responding to post hoc requests, and deciding how much documentation should accompany published results. The talk will also address the responsible use of AI tools in statistical practice. AI may support writing, coding, literature work, exploring new methodology, consistency checks, and documentation, but it does not replace professional judgment or accountability. Its use raises important questions about verification, confidentiality, reproducibility, transparency, and disclosure. The central message is that professional biostatistics is not only about applying appropriate methods, but also about maintaining a transparent and defensible chain from data to interpretation - including any AI-assisted steps.

Early Career Researcher Day Session 3

Thu, 01.10.26, 13:30–15:00, Room: K5

Chair: Autumn Johnson, Max Behrens, Judith Wodke

Session 3

13:30-14:30 Interactive Session

14:30-15:00 Panel Discussion - Voices from Early Career Researchers

Session Descriptions Speed Networking: Attendees are divided into groups of about ten. Each participant has two to three minutes to give an elevator pitch or three-minute thesis style introduction of themselves and their work or research. A great opportunity to showcase your work while making new connections.

Dilemma Game: Building on the success of last year's session, this year's ECR Day includes an interactive session where all participants discuss the challenges statisticians face in conducting biostatistical research. We use the Dilemma Game app, developed by Erasmus University Rotterdam (EUR),

to pose questions and scenarios related to professionalism and integrity in research for discussion. We recommend that all attendees download the app before the mini-symposium.

Panel Discussion - Voices from Early Career Researchers: A key aim of the ECR Day is to provide a forum for discussing the challenges biostatisticians face. Rather than the traditional format where established researchers share their experiences with a largely listening audience, this session gives early-career researchers an opportunity to discuss the opportunities, challenges, and realities of building a career in statistics and biostatistics today. Topics may include research and collaboration, publishing, career development, balancing competing demands, and navigating transitions between academia, industry, and healthcare. The session aims to encourage open discussion, shared experiences, and peer learning, reflecting the collaborative spirit of ECR Day.

We invite all delegates registered for ECR Day who identify as early career researchers to put themselves forward to join the panel. A range of perspectives and backgrounds is welcome, and no previous panel experience is required. Please register your interest here: https://forms.office.com/e/L3csf-pyDHL

SC/AG Meetings

Meeting Details - sorted by date and time

GMDS Präsidiumssitzung

Sun, 27.09.26, 10:00–15:30, Room: N2

Organizer: Annalena Herzog

ISCB ExCom meeting

Sun, 27.09.26, 11:00–18:00, Room: N5

Organizer: Ioanna Aslanidou

GMDS Beiratssitzung

Sun, 27.09.26, 16:00–18:30, Room: N2

Organizer: Annalena Herzog

Empfang GMDS Neumitglieder

Mon, 28.09.26, 12:30–13:15, Room: K1

Organizer: Annalena Herzog

GMDS AG Consumer Health Informatics (CHI)

Mon, 28.09.26, 12:30–13:15, Room: K4

Organizer: Veronika Strotbaum

GMDS AG Health Technology Assessment (HTA)

Mon, 28.09.26, 12:30–13:15, Room: N4

Organizer: Kirsten Herrmann

GMDS AG Management von Informationssystemen im Krankenhaus (MIK)

Mon, 28.09.26, 12:30–13:15, Room: N2

Organizer: Franziska Jahn

GMDS AG MI-Lehre in der Medizin

Mon, 28.09.26, 12:30–13:15, Room: N5

Organizer: Julian Varghese

GMDS Fachausschuss Epidemiologie

Mon, 28.09.26, 12:30–13:15, Room: N1

Organizer: Nicole Rübsamen

GMDS Präsidiumskommission Ethische Fragen in der Medizinischen Informatik, Biometrie und Epidemiologie

Mon, 28.09.26, 12:30–13:15, Room: K3

Organizer: Nils Freyer

GMDS Präsidiumskommission Zertifikatskommission Biometrie in der Medizin

Mon, 28.09.26, 12:30–13:15, Room: K5

Organizer: Antonia Zapf

ISCB Student Conference Awards Committee

Mon, 28.09.26, 12:30–13:15, Room: K2

Organizer: Ioanna Aslanidou

Empfang GMDS-Neumitglieder

Mon, 28.09.26, 12:45–13:15, Room: K1

Organizer: Annalena Herzog

AGM GMDS

Mon, 28.09.26, 17:00–18:30, Room: Rolf Böhme Saal

Organizer: GMDS Board

GMDS Mitgliederversammlung

Mon, 28.09.26, 17:00–18:30, Room: Runder Saal

Organizer: Annalena Herzog

GMDS Präsidiumskommission Zertifikatskommission Medizinische Dokumentation

Mon, 28.09.26, 17:00–18:30, Room: N4

Organizer: Claudia Ose

GMDS & IBS AG Lehre und Didaktik der Biometrie

Tue, 29.09.26, 12:30–13:15, Room: K5

Organizer: Ursula Berger

GMDS AG Curricula der Medizinischen Informatik (CMI)

Tue, 29.09.26, 12:30–13:15, Room: N2

Organizer: Oliver J. Bott

GMDS AG Entscheidungsunterstützung im Gesundheitswesen

Tue, 29.09.26, 12:30–13:15, Room: K1

Organizer: Zully Ritter

GMDS AG Informationsverarbeitung in der Pflege

Tue, 29.09.26, 12:30–13:15, Room: K3

Organizer: Björn Sellemann

GMDS AG Medizinische Terminologien und Klassifikationen (MTK)

Tue, 29.09.26, 12:30–13:15, Room: N5

Organizer: Julian Saß

GMDS AG Statistische Methodik in der klinischen Forschung

Tue, 29.09.26, 12:30–13:15, Room: K4

Organizer: Verena S Hoffmann

GMDS Fachausschuss Medizinische Biometrie

Tue, 29.09.26, 12:30–13:15, Room: N3

Organizer: Anika Großhennig

GMDS Projektgruppe Diversität

Tue, 29.09.26, 12:30–13:15, Room: N1

Organizer: Michael Rusongoza Muzoora

ISCB Education Subcommittee

Tue, 29.09.26, 12:30–13:15, Room: K2

Organizer: Ioanna Aslanidou

ISCB SC Statistics in Regulatory Affairs (SiRA)

Tue, 29.09.26, 12:30–13:15, Room: N4

Organizer: Nicole Close

AGM ISCB

Wed, 30.09.26, 12:30–13:40, Room: Runder Saal

Organizer: ISCB Executive Committee

GMDS AG Epidemiologie der Herz-Kreislauf- und Stoffwechsel-Erkrankungen

Wed, 30.09.26, 12:30–13:15, Room: N1

Organizer: Sara Schramm

GMDS AG Informationssysteme im Gesundheitswesen KIS

Wed, 30.09.26, 12:30–13:15, Room: N4

Organizer: Viola Henke

GMDS AG Medizinische Bild- und Signalverarbeitung (MBSV)

Wed, 30.09.26, 12:30–13:15, Room: N2

Organizer: Dagmar Krefting

GMDS AG Mobile Informationstechnologie in der Medizin und Telemedizin (MoCoMed)

Wed, 30.09.26, 12:30–13:15, Room: N3

Organizer: Andreas Koop

GMDS AG Therapeutische Forschung

Wed, 30.09.26, 12:30–13:15, Room: K4

Organizer: Anika Großhennig

GMDS Arbeitskreis Epidemiologie der Arbeitswelt

Wed, 30.09.26, 12:30–13:15, Room: K5

Organizer: Benjamin Kendzia

GMDS Fachausschuss Medizinische Bioinformatik und Systembiologie

Wed, 30.09.26, 12:30–13:15, Room: N5

Organizer: Klaus Jung

GMDS Projekt EVA4MII

Wed, 30.09.26, 12:30–13:15, Room: K2

Organizer: Michelle Pfaffenlehner

GMDS Projektgruppe Global Health Informatics (GHI)

Wed, 30.09.26, 12:30–13:15, Room: K1

Organizer: Fleur Fritz-Kebede

Author Index

Aas, Kjersti, 567
Abdelmoula, Adem, 347
Abdulnazar, Akhila, 197
Abecassis, Judith, 1022
Abed, Amal, 349
Abert, Helke, 923
Abootalebi, Amirmahdi, 249
Abrahamowicz, Michal, 220, 1090, 1091
Abrams, Keith, 354, 715, 996
Abrams, Keith R., 204
Abubakar, Amina, 876
Adam, Helena, 197
Adamoli, Arianna, 438
Adams, Robert, 673
Adamski, Jerzy, 891
Adamzik, Michael, 566
Adderley, Nicola, 642
Adeleke, Ireoluwa, 869
Adewuyi, Emmanuel, 431
Agha-Mir-Salim, Louis, 539
Aghdassi, Seven Johannes Sam, 586
Agostini, Andrea, 609
Ahmadi, Najia, 307
Ahmed, Shamim, 942
Ahrens, Maike, 372
Aittokallio, Tero, 266
Akicier, Burak, 378, 725
Akpinar, Meltem Gülsün, 588
Akyianu, Gyamfi, 874
Akynkozhayev, Birzhan, 590
Alam, Moudud, 510
Alanya, Ahu, 1099
Alauddin, Maliha Binte, 688
Albou, Laurent-Philippe, 698
Alcalá, José Tomás, 647
Alderman, Joseph, 642
Alfò, Marco, 848
Ali, Mian Haider, 649
Alickovic, Fatma, 245
Allassonnière, Stéphanie, 706
Allen, Elizabeth, 961
Allenbach, Yves, 1003
Alnahas, Hani, 1069
Alongi, Alessandra, 641
Altenbuchinger, Michael, 304
Alvarez, Nuria Perez, 216
Alwie, Yasmine, 541
Amaral, Andre, 827
Ambler, Gareth, 268, 463, 645, 966
Ambrogi, Federico, 438, 1096
Amft, Oliver, 225, 355, 407
Ammer-Herrmenau, Christoph, 778
Ammon, Danny, 541, 613, 618, 619
Amstutz, Alain, 94
Anand, Prateek, 541
Anastasova, Ivona, 605, 624
Anderle, Laura, 703
Andersen, Per Kragh, 507
Andersen, Tobias Søren, 440
Anderson, Ilona, 975
Andersson, Therese, 339, 508, 978
André, Pascaline, 299
Andrinopoulou, Eleni Rosalina, 852
Andueza, Mireia, 492, 647
Angleitner, Alexander Casimir, 541
Ansari, Aseer, 748
Antolini, Laura, 992, 994
Antoniou, Antonis, 661
Apalategui, Pedro Octavio Markus Gebhard, 190
Apeldoorn, Daan, 487, 664
Apfel-Starke, Johanna, 277, 615, 616
Apostolidou-Kiouti, Fani, 734
Appel, Katharina, 803
Arai, Takayuki, 696
Archambaud, Baptiste, 828
Archer, Lucinda, 642, 643

Arendt, Marina, 703
Arjas, Elja, 213
Armand, Stéphane, 840
Arnold, Philipp, 1069, 1077
Arvandi, Marjan, 328, 699, 837, 1023
Arvedsen, Sine Kongsbak, 440
Asano, Riko, 674
Ashok, Amra Amra Dhabalia, 633
Assaad, Charles, 435
Attieh, Hammam Abu, 100, 101, 102, 103
Aubele, Fabio, 484
Audebert, Heinrich, 943
Augusto, Margarida, 386
Auranen, Kari, 477
Austin, Peter C., 341
Avci, Hanife, 588
Awad, Noor, 1074
Awounvo, Sinclair, 650
Azevedo, Sofia, 379, 960
Azócar, Antonio Remiro, 386
Bøg, Martin, 386
Bıyıklı, Furkan, 935
Baas, Stef, 427
Bacak, Miroslav, 864, 987
Bach, Paul, 879
Bachir, Sara, 537
Backofen, Rolf, 1047, 1057, 1064, 1065
Bacry, Emmanuel, 289
Badisy, Imad El, 651
Badpa, Mahnaz, 545
Baghfalaki, Taban, 291, 720
Bahmer, Thomas, 803
Baillie, Mark, 958
Baitairian, Jean-Baptiste, 482
Bakewell, Nicholas, 1093
Bakony, Mikolt, 739
Bakoyannis, Giorgos, 515
Balci, Yusufhan, 534
Baldomero, Helen, 611
Ballegaard, Martin, 440
Ballhausen-Lübcker, Sabrina, 803
Baltus, Hannah, 1005
Balzer, Alexandra, 675
Balzer, Felix, 147, 159, 176, 178, 409, 539, 806, 931
Bamberg, Fabian, 891, 1017
Bambury, Pauline, 960
Ban, Hideyuki, 696, 796, 949
Bandak, Elisabeth, 986
Banerjee, Mousumi, 415, 985
Banfi, Cristina, 883
Bannay, Aurélie, 276
Bannick, Marlena, 958
Bao, Loc Carlo, 519
Barbati, Giulia, 392, 660
Barbet, Vanessa, 1029
Bardram, Jakob Eyvind, 440
Bardtke, Adrian, 191
Barreñada, Lasai, 91, 122, 646
Barrenetxea, Jon, 931
Barrett, Jessica, 201
Bartlett, Jonathan, 431, 1093
Bartram, Malte, 625
Bartsch, Madeline, 189
Baschet, Louise, 578, 1029
Bast, Hannah, 349, 1050, 1069, 1073, 1077
Bastian, Patrick, 85
Baudot, Anaïs, 712
Bauer, Antje, 586
Bauer, Hans, 149, 151
Bauer, Robert, 163, 165, 386
Bauersachs, Johann, 755, 1001
Baumann, Lukas, 522, 523
Baumann, Tristan, 190
Baumeister, Vera, 527
Bavendiek, Udo, 540, 755, 1001
Bayer, Malte, 566
Bean, Andrew, 469
Beauchamp, Marie-Eve, 1090
Bebu, Ionut, 241
Becher, Matthias, 943

Beck, Jördis, 618
Beckmann, Lars, 821
Bedir, Ahmed, 999
Beerenwinkel, Niko, 142, 143, 156
Beernink, Mareile, 233, 947
Behling, Stefan, 409
Behning, Charlotte, 794
Behnke, Michael, 586
Behrend, Christina Birgit, 326
Behrendt, Jordan, 789
Behrens, Max, 417, 420, 1046, 1059, 1061
Behrens, Thomas, 854
Beier, Frank, 261
Beißbarth, Tim, 142, 156
Belle, Fabiën, 609
Belleli, Rossella, 204
Bellerino, Gabriele, 417, 1081
Bellio, Ruggero, 422
Bemelman, Frederike J., 1000
Ben, Tahan, 1018
Bender, Ralf, 163, 164, 383, 731
Beneton, Maud, 960
Benveniste, Olivier, 1003
Benzekry, Sebastien, 717
Benzoni, Claudio, 902
Beppler, Tatjana, 484, 753, 780, 899
Berchialla, Paola, 940
Berchtold, Luzia, 577
Berendonck, Maren, 703
Berens, Philipp, 626
Berg, Daniela, 708
Berger, Felix, 802
Berger, Stephanie, 248, 376, 783, 1036
Berger, Ursula, 184
Bergmann, Lars, 566
Berkhof, Johannes, 502, 811
Berliner, Dominik, 1001
Bernasconi, Davide, 992
Berthe, Sebastian, 626
Best, Nicky, 172, 193
Betschart, Raphael O., 633
Beuermann, Janik, 873
Beyan, Oya, 222, 325, 625
Beyer, Nils, 778, 983
Beyersmann, Jan, 836
Beyrich, Katharina, 1027
Bhaskarapillai, Binukumar, 790
Bhat, Bethany Hamilton, 867
Bialke, Martin, 760, 761, 771, 772
Biard, Lucie, 470, 473, 670
Bicknell, Colin, 436
Bidulka, Patrick, 582, 833
Biedenkapp, André, 1074
Bienzeisler, Jonas, 247
Billinger-Finke, Mareike, 975
Binder, Harald, 124, 255, 349, 352, 417, 420, 494, 495, 705, 794, 1046, 1052, 1059, 1060, 1061, 1062, 1063, 1066, 1071
Binder, Nadine, 124, 352, 1041, 1061, 1066
Binet, Isabelle, 928
Binois, Yannick, 1022
Bittmann, Janina A., 779
Blanchard, Ilona, 717
Blasini, Romina, 618
Blaß, Michael, 233
Bleckmann, Annalen, 580
Bleich, Lene, 190
Blettner, Maria, 979
Blieske, Pio, 839
Bluhmki, Tobias, 163, 165
Blumentritt, Arne, 772
Bobrowski, Christoph, 144, 146
Bodden, Daniel, 562
Bode, Lisa Kunkel genannt, 299
Boedecker, Joschka, 1071, 1074
Boeker, Martin, 196, 231, 232, 347, 375, 902
Böhm, Samuel, 493, 1070
Böhme, Sarah, 163, 165
Bohn, Stefan, 769
Böhringer, Jan, 629
Boie, Sebastian, 409, 806
Boiko, Olga, 1019

Bokern, Marleen, 690
Bonanno, Laura, 519
Bonhoeffer, Sebastian, 839
Bönisch, Caroline, 168, 170
Bönisch, Nicole, 616
Bonneville, Edouard, 412
Booth, Sarah, 467, 845
Borena, Wegene, 593
Borisov, Oleg, 493, 571
Bormann, Eike, 1006
Börnhorst, Claudia, 789
Bornkamp, Björn, 172, 174, 193, 596
Bosi, Emanuele, 1019
Botea, Gabriela, 134
Bothe, Carola, 586
Bott, Oliver J., 441, 604
Bottai, Matteo, 978
Bouee, Stéphane, 742
Boulesteix, Anne-Laure, 186, 235, 411, 412, 686, 722, 926
Boulet, Sandrine, 488, 489, 501, 828, 1022
Boumendil, Luana, 364
Bousema, Teun, 320
Boutillier, Stéphanie, 740
Bové, Daniel Sabanés, 203, 958
Bozek, Katarzyna, 142, 156
Bozorgmehr, Kayvan, 770
Bozorgtabar, Behzad, 1069
Bracher, Johannes, 107
Bracht, Thilo, 566
Bradley, John, 212
Brambilla, Marta, 883
Brandstetter, Lilly Sophia, 223
Brandt-Heunemann, Stephanie, 606, 608
Brannath, Werner, 309, 363, 513
Brathovde, Mari, 476
Bratulić, Jelena, 1049, 1050, 1068, 1072, 1073
Brauer, Susan, 922, 923
Braun, Annika, 223
Bräutigam, Jacob, 638
Breger, Max, 247
Breidenbach, Jan Sören, 190
Breil, Bernhard, 916
Breitfeld, Philipp, 223
Bretz, Frank, 596, 722
Bricullé, Pau Esteve, 216
Briel, Matthias, 94, 411, 722, 955, 1098
Brigiari, Gloria, 381, 519, 788, 842, 860, 1008
Brix, Tobias, 326, 580, 626, 628, 773, 900
Brockmann, Kathrin, 708
Broeder, A.A. den, 405
Brookfield, Melissa, 877
Brors, Benedikt, 799
Brox, Thomas, 1049, 1050, 1052, 1060, 1068, 1069, 1072, 1073
Bruder, Nico, 562
Brüggen, Heleen, 951
Brugger, Silvio, 839
Brülhart, Damian, 655
Brunn, Niklas, 495, 1047, 1065
Brüx, Hannah, 190
Buch, Gregor, 370, 384
Buchan, Iain, 835
Buchholz, Simone, 663
Buchner, Hannes, 149, 926
Bücken, Gerrit, 299
Buckow, Karoline, 618
Budde, Klemens, 225
Budeus, Markus, 232
Bujkiewicz, Sylwia, 869
Bülüç, Hüseyin, 588
Bulut, Nurgül, 740
Bumbulis, Laura, 553
Burges, Ciel, 198
Burhan, Hüseyin Şehit, 935
Burnell, Matthew, 702
Busch, Luca, 191
Büscher, Sarah, 916
Büscher, Sebastian, 413
Butz, Elena, 794
Buyze, Jozefien, 927
Bystrova, Daria, 435

Cadarette, Suzanne, 1093, 1097
Cadi, Loubna, 489
Caille, Agnes, 212
Cakir, Enes Malik, 584
Caldera, Luca, 294, 337
Calderazzo, Silvia, 172, 193
Caliebe, Amke, 548
Calik, Emre, 896
Callegaro, Andrea, 172, 173, 193, 500
Camanzi, Barbara, 1031
Camera, Marina, 883
Canas, Liane, 1009
Cao, Charlie, 811
Cao, Zhi, 524
Capitoli, Giulia, 497, 641, 889
Cappozzo, Andrea, 337
Carlin, John B., 75, 433
Carlini, Luca, 218
Carneado, Jimena Fernández, 216
Carpenter, James, 84, 124, 215, 702, 1090, 1094
Carr, Ewan, 227, 287, 956
Carriero, Alex, 391
Cartailler, Jérôme, 1029
Carter, Ben, 198
Casimiro, Gabrielle, 706
Cassese, Alberto, 546
Castagné, Claire, 437, 698
Castel, Charlotte, 831
Castelletti, Noemi, 932
Castro, César Andrés Pérez, 235
Cate, Vincent ten, 370, 875
Cathomen, Toni, 1042, 1064
Catoire, Pierre, 265
Catozzi, Simona, 960
Causeur, David, 861
Cavallaro, Alexander, 763
Cavinato, Lara, 294, 556
Cenko, Egidia, 485
Centanni, Maddalena, 529
Cesa, Alessandro, 768
Cevirme, Sinan Necdet, 228, 726
Chaimani, Anna, 333
Chakraborty, Hrishikesh, 558
Chalkou, Konstantina, 385, 503, 870
Chambaz, Antoine, 573
Chammartin-Basnet, Frédérique, 964
Chaplinskaya-Sobol, Irina, 803
Charalambous, Christiana, 810
Chauveau, Bertrand, 1000
Chelangat, Daisy, 876
Chen, Dung-Tsa, 965
Chen, Sida, 565
Chen, Yuxuan, 894
Chen, Zerong, 827
Chen, Zhaojin, 963
Chereda, Hryhorii, 745
Cherlin, Svetlana, 1032
Chesang, Caroline, 436
Chester-Jones, Mae, 649
Chevret, Sylvie, 470, 473, 549, 670, 829, 850
Chiaborelli, Malena, 411, 722
Chiocchia, Virginia, 279
Chiriac, Ute, 398
Choodari-Oskooei, Babak, 314, 680, 682
Chowdhury, Vivek, 959
Christodoulou, Maria D, 988
Christoffersen, Benjamin, 590
Christoph, Jan, 244, 346, 799, 908, 909
Chu, Petrina, 198
Chudek, Jerzy, 1018
Ciardulli, Simone, 692
Cieśla, Daniel, 777
Cipriani, Andrea, 279
Cirici, Marc Olivella, 1023
Cizauskas, Nicole, 1032
Clairon, Quentin, 251
Clements, Mark, 590
Clever, Lena, 580
Close, Nicole, 377, 822, 968
Coates, Laura, 405

Coelho, Vasco, 497
Coemans, Maarten, 1000
Coffeng, Luc, 320
Coll-Planas, Laura, 699, 837
Collignon, Corinne, 1022
Collins, Gary, 89, 391, 462, 642, 643, 649, 877, 1095, 1100
Collyer, Taya, 412
Columbus, Alyssa, 235
Comets, Emmanuelle, 472, 717
Connell, David Mc, 163, 166
Connolly, Patrick F., 975
consortium, Julia Bielicki on behalf of SwissPed-Health, 609
Constantine-Cooke, Nathan, 721
Conti, Valentino, 500
Conyers, Rachel, 153
Cook, Richard, 553
Cooper, Nicola, 733
Cordes, Steffen, 605
Cornish, Rosie, 215
Cortese, Giuliana, 509
Coser, Andrea, 1030
Cossio, Eliana Garcia, 157, 372
Costagliola, Dominique, 94
Crespo, Leandro Teixeira, 664
Cro, Suzie, 361
Crooks, Colin, 974
Cross, Samantha, 198
Cuijpers, Pim, 1084
Culhane, Aedin, 1007
Curnow, Elinor, 215, 687
Cypko, Mario A., 225, 355
D'Errigo, Paola, 694
Dählmann, Klaas, 1028
Dai, Bingying, 731
Dallavalle, Marco, 637
Damm, Hendrik, 724
Dammer, Ulf, 971
Dandl, Susanne, 531
Danhauser, Katharina, 147, 159, 176, 178, 179
Danzer, Moritz Fabian, 817
Das, Rajenki, 563
Dashti, S Ghazaleh, 75
David, Alexandra, 367
Davies, Gwyneth, 262
Davies, Mark, 354
De Bin, Riccardo, 266, 1096
de Hond, Anne, 391
de Jong, Valentijn, 602, 735
de Leeuw, Robert, 434
De Marco, Gabriele, 405
de Maturana, Evangelina López, 492
De Pourcq, Sven, 144, 145
De Santis, Riccardo, 422
de Uña-Álvarez, Jacobo, 239
de Vries, Aiko P.J., 1000
de Wreede, Liesbeth, 107, 110, 292, 318, 552, 658
Debebe, Michael, 406
Debyser, Tim, 1000
Dedié, Angela, 605, 624
Deep, Aman, 897
Deffner, Veronika, 980
Degan, Chiara, 707
Dekker, Andre, 69
Del Castello, Lorenzo, 992
Delgado, Inigo Bermejo, 876
Deligianni, Xeni, 609
Delvaux, Bernard, 1029
Delvo, Agatha Sanchez, 919
DEMİRKAZIK, Figen, 588
Demus, Christoph, 244, 346, 908, 909
Denecke, Kerstin, 730
Dennis, Brittany B., 871
Denniston, Alastair, 642
Dennstädt, Fabio, 1035
Denti, Vanna, 497, 889
Denz, Robin, 695, 832
Derad, Carlotta, 398
Derici, Didem, 335
Detering, Bianka, 770
Dette, Holger, 85, 596, 1051, 1058, 1082

Dettmer, Daniel, 761
Devane, Declan, 94, 96
Dhiman, Paula, 642, 877
Dhungana, Hom Nath, 316, 322
Di Angelantonio, Emanuele, 446, 581
Di Mari, Fabrizio, 848
Di Mattei, Valentina, 1019
Di Ran, 787
Di Serio, Clelia, 1019
Di Stefano, Leon, 152, 153, 455, 677, 824
Diao, Guoqing, 241
Diaz, Luis Alberto Peña, 586
Diaz, Mireya, 1035
Diaz-Ordaz, Karla, 76, 582
Dickhoff, Lothar, 626
Didelez, Vanessa, 80
Dieci, Maria Vittoria, 381
Diekmeier, Oliver, 625
Dienert, Johannes, 1050, 1073
Dierkes, Jens, 325
Dieterich, Christoph, 896, 906
Dietscher, Christina, 593
Dilje, Marcel, 614
Ding, Yuwen, 318
Ding, Ziyi, 394
Diop, Awa, 517
Diouf, Matylde, 829
Dittmann, Kathleen, 772
Dittrich, Henriette, 979
Dobler, Valeriia, 1011
Dobre, Dragos, 486
Dobre, Simona, 486
Dolouei, Maryam Khodaei, 769
Don, Lise, 703
Dong, Trinh, 358, 1009
Donovan, Joseph, 1009
Dormann, Carsten F., 1071, 1074
Downing, Jennifer, 659
Draeger, Christian, 230, 396
Drechsler, Matthias, 759
Dreher, Matthias, 664
Drepper, Johannes, 770
Drescher, Lennart, 634
Drescher, Sonja, 181, 182
Driusso, Eugenia, 267
du Montcel, Sophie Tezenas, 576, 706
Duda, Antje, 580
Düking, Peter, 634
Dumas, Elise, 531
Dümmler, Daniel, 544, 1011
Dumontier, Michel, 71
Dumville, Jo, 851
Dunkler, Daniela, 228, 235, 726, 1098
Durhan, Gamze, 588
Duwe, Gregor, 350
Easton, Douglas, 661
Eberhardt, Martin, 191
Ebertsch, Kathrin, 765
Ebigbo, Alanna, 147, 148, 159, 176, 178
Eckardt, Kai-Uwe, 794
Eckert, Alexander, 624
Edel, Andreas, 806
Edelmann, Dominic, 478
Edemir, Bayram, 493
Eden, Arzu Baygul, 335
Ederhy, Stephane, 752
Efthimiou, Orestis, 450, 503, 1084
Egbert, Nicole, 327
Eibensteiner, Fabian, 181, 183
Eichhorst, Barbara, 222
Eil, Jan, 538
Eisele-Metzger, Angelika, 882
Eisemann, Nora, 1005
Eisenacher, Martin, 566
Eisfeld, Leo-Lion, 900
El-Helou, Salhab, 1026
Elayan, Haya, 394
Elefsinioti, Antigoni, 157, 372
Eletti, Alessia, 500
Elgert, Lena, 604

Ell, Basil, 266
Ellenberger, David, 585
Ellmann, Tanja, 700
Elmaleh, Yoann, 1029
Elmer, Andreas, 928
Elsaesser, Moritz, 1084
Elsbernd, Kira, 932
Elze, Markus, 698, 819
Emery, Corinne, 742
Emili, Luca, 249, 379, 768
Emoto, Ryo, 662, 747
Emsley, Richard, 198
Enders, Dominic, 791
Endres, Matthias, 943
Engelbrecht, Alexander, 614
Engelemann, Katrin, 621
Engelhardt, Lilian J., 806
Englund, Martin, 986
Ensor, Joie, 91, 462, 642, 643, 644
Ensslen, Tobias, 407
Enste, Peter, 703
Erdmann, Stella, 728, 846
Eriksson, Frank, 401
Eriksson, Hanna, 978
Erkorkmaz, Unal, 884
Erler, Nicole, 202, 218, 267, 391
Ertaylan, Gokhan, 168, 169
Escanciano, Juan Carlos, 239
Eschweiler, Gerhard, 708
Eskelson, Louisa, 191
Eskofier, Bjoern, 147, 159, 176, 178
Esteban, Luis Mariano, 647, 809
Esteban-Escaño, Javier, 809
Eter, Nicole, 900
Evrenoglou, Theodoros, 331, 334, 417, 1080, 1084
Fahy, Martin, 1021
Faldum, Andreas, 997
Fallgatter, Andreas, 708
Farhadizadeh, Maryam, 352, 1061, 1066
Farhadyar, Kiana, 495, 1042
Farin-Glattacker, Erik, 882, 914
Fässler, Daniel, 1075
Faujour, Corentin, 742
Federbusch, Martin, 324, 484
Federmann, Lydia, 708
Fegeler, Christian, 629, 783
Fehr, Thomas, 928
Feißt, Manuel, 133, 650
Felderhoff, Hendrik, 626
Fellinger, Tobias, 183, 529
Felsch, Moritz, 821
Fenske, Nora, 980
Ferfache, Anouar Abdeldjaoued, 242
Ferreira, Simon, 435
Fette, Georg, 540
Feydy, Jean, 289, 437
Fichter, Carola, 882
Fichtner, Urs Alexander, 882, 914
Ficorella, Lorenzo, 661
Fiedler-Lacombe, Lizon, 772
Fierenz, Alexander, 544, 545
Filleron, Thomas, 578
Fiocco, Marta, 317, 480, 656, 864
Firat, Huseyin, 740
Fischer, Lasse, 363
Fischer-Betz, Rebecca, 700
Fischhuber, Karen, 1006
Fisher, David, 599
Fitzer, Kai, 445
Fitzgerald, Niamh, 521
Flake, Julian, 230
Flamigni, Alice, 253
Fleck, Christian, 1049, 1072
Fleischmann, Maximilian, 541
Fleßner, Jannik, 604
Flis, Martyna, 1017
Florquin, Sandrine, 1000
Fneish, Firas, 585
Foco, Luisa, 415
Folwark, Julia, 799

Folz, Michael, 537
Fontaine, Guillaume, 561
Fontana, Nicole, 253, 692
Fontana, Saverio, 731
Foraita, Ronja, 397
Forbes, Andrew, 560
Forbes, Gordon, 198, 227, 287, 956
Fortwengel, Gerhard, 585
Foubert, Alexandra Samier, 358
Foubert-Samier, Alexandra, 205
Fox, Emily, 682
Frahm, Niklas, 585
Frangou, Elena, 314
Frank, Derk, 803
Frank, Julia, 823
Franke, Georg-Nikolaus, 541
Franscini, Nicola, 928
Frase, Sibylle, 705
Freedman, Laurence, 220, 1091
Freeman, Carol, 436
Freeman, Suzanne, 733
Freiesleben, Sherry, 396, 445
Freund, Robert, 540
Frey, Christian, 1083
Frey, Nicolas, 409
Friebe, Sabrina, 345
Friede, Tim, 133, 172, 181, 182, 193, 311, 529, 584, 819, 874
Friedrich, Christoph M., 724
Friedrich-Welz, Sarah, 238
Fritz, Adrian, 1047, 1057, 1065
Fritz, Laetitia, 237
Frömke, Cornelia, 755, 1001
Froněk, Jiří, 1002
Fröschl, Barbara, 593
Fryc, Jaga, 877
Fu, Han, 541
Fuchs, Timo, 762
Fudickar, Sebastian, 134
Fuellen, Georg, 144, 145
Fukai, Keita, 990
Fullio, Camila L., 495
Funderich, Meike T., 605
Furberg, Julie Funch, 386
Furukawa, Toshi A., 1084
Fusch, Gerhard, 1026
Fusiak, Jakub, 917
Gademan, Maaike, 786
Gadhia, Nandini, 1031
Gaertner, Thomas, 924
Gail, Mitchell H, 1097
Galimberti, Stefania, 365, 641, 710
Gall, Christine, 971
Gallizzi, Elias, 922, 923
Galotta, Arianna, 883, 1030
Gambazza, Simone, 438
Ganjali, Mojtaba, 720
Ganslandt, Thomas, 114, 116, 127, 485, 762, 763
Ganzinger, Matthias, 612
Garabetian, Garo, 332
Garcelon, Nicolas, 276
Garcia-Fernández, Arnau, 343
Garner, Malvina, 748
Gartlehner, Gerald, 593
Gärtner, Thomas, 217
Gashi, Bora, 231
Gąsior, Mariusz, 777
Gasparini, Alessandro, 203
Gatrio, Margaux, 396
Gatti, Francesca, 1019
Gatz, Susanne A, 801
Gatzemeier, Hanna Elena, 905
Gauges, Sarah, 140
Gaunt, Piers, 801
Gehrig, Stefan, 879
Geler, Iris, 920
Génin, Emmanuelle, 712
Genuer, Robin, 265
George, Sneha, 835
Gera, Roland Gerard, 398
Gérard, Emilie, 1021

Gerbach, Sarah, 920
Gerds, Thomas Alexander, 574
Gerhold-Ay, Aslihan, 131
Geroldinger, Angelika, 529
Gerry, Stephen, 649
Gershon, Richard, 985
Geršlová, Zdeňka, 911
Gervytė, Miglė, 243
Gerwek, Marie, 925
Geskus, Ronald, 1009
Geyer, Christina, 737
Gharibzadeh, Safoora, 844, 989
Ghestem, Florence, 712
Gholam, Mehdi, 714
Ghorbani, Alireza, 412
Gianicolo, Emilio, 685, 979
Giannoni, Eric, 609
Giardiello, Daniele, 464
Giesa, Niklas, 409
Giese, Robin Max, 632
Giesinger, Johannes, 982
Gieswinkel, Alexander, 370
Gillies, Clare, 844
Ging, Simon, 1049, 1050, 1069, 1072, 1073
Ginnerup-Nielsen, Elisabeth, 986
Giordano, Sabrina, 848
Giorgi, Roch, 242, 651
Girdziunaite, Gerda, 538
Gitiaux, Cyril, 1003
Giudice, Enrico, 575
Gkatzionis, Apostolos, 400, 533, 687
Glattacker, Manuela, 920
Glemain, Benjamin, 435
Glen, Callum, 198
Glimm, Ekkehard, 362
Glöckner, Stefan, 619
Gnimatin, Jean-Pierre, 784
Göbel, Jens, 273, 774, 775
Göbel, Nicole, 609
Godec, Aljaž, 1081
Godfrin, Yann, 960
Goedhart, Jeroen, 567
Goeman, Jelle, 426
Goetghebeur, Els, 206, 1094, 1099
Gögele, Martin, 415, 985
Goksuluk, Dincer, 884
Goksuluk, Merve Basol, 884
Gokulakrishnan, Kuppan, 790
Golchi, Shirin, 561
Golebiewski, Martin, 168, 640
Golla, Heidrun, 398
Golshayan, Déla, 928
Gómez, Miriam Pedrera, 459
Gomon, Georgy, 551
Gonzato, Elia, 502, 811
Goodall, Mark, 659
Göpel, Wolfgang, 973
Gordon, Susanna, 196
Gorlia, Thierry, 202
Görlich, Dennis, 997, 1006
Götte, Heiko, 131, 261
Grabocka, Josif, 1083
Graf, Erika, 892
Graf, Lennart, 194
Grafe, Vera, 873
Graffeo, Nathalie, 242
Grandelis, Elena, 975
Granger, Emily, 262
Grashof, Robin, 112, 916
Grathwohl, Dominik, 954
Gravestock, Isaac, 459
Graziano, Francesca, 710
Green, Nathan, 645
Gregori, Dario, 381, 519, 788, 842, 860, 1008
Greve, Gabriele, 493
Grieger, Paula, 1005
Grieve, Andy, 172, 193
Griewank, Klaus, 792
Griguola, Tomas, 243
Grimm, Ulfert, 924

Grittner, Ulrike, 943
Gröber, Petra, 445
Grobler, Anneke, 153
Groll, Andreas, 238
Gronwald, Wolfram, 794
Gross, Christopher, 922, 923
Gross, Franziska, 982
Große, Silke, 905
Grosse-Meininghaus, Dirk, 540
Großhennig, Anika, 131, 149, 163, 259, 755, 1001
Groten, Tanja, 905, 930
Grouven, Ulrich, 383, 821
Gruber, Raphael, 924
Gruendner, Julian, 536
Grugel, Robin, 566
Grummt, Sophia, 621
Grune, Elena, 686, 891
Grünewald, Elias, 409, 539, 806, 931
Grünwald, Peter, 426
Grzegorzek, Marcin, 299
Gu, Tianyuan, 210, 963
Guarneri, Valentina, 381
Gubelli, Gaia, 889
Guerra, Steve Ferreira, 220, 1091
Guessous, Karim, 1029
Guillon, Antoine, 472
Guilloux, Agathe, 254, 482, 516
Guitart, Laia C., 495
Gül, Gökhan, 370, 875
Gundler, Christopher, 233, 947
Gunel, Pınar, 740
Günther, Andreas, 904
Guo, Linlin, 633
Guolo, Annamaria, 422
Gupta, Ayush, 1068
Gürbulak, Selin, 308
Gustafson, Paul, 220, 1091
Guyatt, Gordon, 722
Haase, Isabell, 700
Haben, Sabrina, 421
Habenicht, Hannah, 1061, 1071, 1074
Habermehl, Christina, 261
Hackenberg, Maren, 306, 494, 495, 1047, 1061, 1063, 1065
Hackl, Monika, 593
Hackmann, Toby, 292, 390, 658
Hadaschik, Eva, 792
Hadjur, Hugo, 372
Haeusler, Karl Georg, 803
Hafermann, Lorena, 217
Hagemeier, Anna, 823
Hägglund, Gunnar, 805
Hähner, Felicia, 608
Haidar, Fadi, 928
Haidich, Anna-Bettina, 332, 734
Hakova, Tereza, 1002
Halbritter, Jan, 625
Halilovic, Mehmed, 100, 101, 102, 103
Hallsson, Lára R., 328, 593, 699, 837
Haltiner, Rafaela, 825
Hamano, Jun, 998
Hampf, Christopher, 626, 760, 761
Hamza, Tasnim, 204, 385
Han, Shawn, 201
Han, Youmei, 460
Handels, Heinz, 744
Hanif, Umaer Rashid, 440
Hansen, Clint, 708
Hansmann, Helena, 794
Hanß, Sabine, 273, 443, 540, 606, 608, 757
Hanßmann, Franziska, 973
Hanßum, Luzia, 926
Hapfelmeier, Alexander, 280, 282
Hargreaves, Suzannah, 659
Härkänen, Tommi, 477
Harms, Bendix, 190
Harrel, Frank E., 220, 341, 1091
Hartmann, Michaela, 570
Hartmann-Boyce, Jamie, 733
Hartung, Lisa, 370, 875
Hartwig, Fernando, 533

Harvey, Joe, 851
Has, Cristina, 257, 704, 1048, 1068, 1076, 1078
Hasan, M Asif, 1074
Hasegawa, Yasutaka, 696, 949
Hasenauer, Jan, 352, 1041, 1063, 1066
Hashimoto, Atsuki, 260
Hattori, Satoshi, 423, 424, 506, 525
Haucke, Marie, 272
Haug, Nina, 691, 834
Haug, Stefan, 493
Hauschild, Anne-Christin, 199, 514, 745
Haverberg, Maria Joy Normann, 440
Hebestreit, Nicolas, 190
Hecht, Manuel, 305
Hechtel, Nicole, 616
Hedberg, Jonatan, 508
Heesen, Philip, 499, 666
Heesterman, Berdine, 292
Hegenauer, Joachim, 663
Heidel, Andrew, 613, 619
Heiden, Pierce, 823
Heidenreich, Andreas, 537
Heider, Dominik, 580, 620, 628, 773, 900
Heim, Ekatarina, 772
Hein, Andreas, 1028
Heinemann, Stephanie, 398
Heines, David, 626
Heinrich, Philipp, 760
Heinz, Sinah, 764
Heinze, Georg, 228, 234, 286, 411, 412, 464, 1093, 1098
Heinzel, Carola S., 352, 1051, 1058, 1082
Heinzel, Sebastian, 708
Heitz, Charles, 299
Hejblum, Boris, 569
Held, Leonhard, 731
Held, Ulrike, 655, 825
Heller, Gillian, 858
Hellmich, Martin, 181, 182, 398, 823
Hellwig, Birte, 792
Hellwig, Dirk, 749
Hellwig, Julius, 190
Hellwig, Kerstin, 421
Helm, Kathrin, 920
Hemkens, Lars G, 955
Hempel, Philip, 168, 171, 745
Hempler, Isabelle, 920
Henke, Christian, 230, 632
Hennigs, Nadine, 566
Henriksen, Marius, 986
Henrion, Marc, 1098
Hentschel, Tino, 922, 923
Hering, Annika, 323, 327
Hermann, Tobias, 782
Heron, Jon, 215
Herr, Marius, 100, 101, 102, 103
Herrmann, Carolin, 184, 209, 302, 652, 668, 1033
Herrmann, Jochen, 744
Herrmann, Kirsten, 104
Herrmann-Lingen, Christoph, 398
Hertel, Johannes, 1044, 1061, 1075
Herzig, Anthony, 712
Hess, Moritz, 257, 417, 704, 1048, 1076, 1078, 1085
Hessmann, Elisabeth, 983
Hettle, Robert, 787
Heuschmann, Peter, 803
Heuschmann, Peter U., 223
Heuser, Michael, 541
Heyckendorf, Jan, 803
Hida, Eisuke, 816, 818, 830
Hieber, Daniel, 589
Hieber, David, 100, 101, 102, 103
Hilgendorf, Inken, 541, 1005
Hilgers, Ralf-Dieter, 360, 562
Hillemanns, Peter, 593
Hindel, Julia, 1068
Hinske, Christian, 147, 159, 176, 178
Hinske, Ludwig Christian, 764, 802
Hirt, Julian, 411, 722
Hochstrat, Beate, 980
Hocke, Sebastian, 748

Hodsoll, John, 956
Hoessly, Linard, 928
Hofenbitzer, Justin, 375
Hoffmann, Justin Steven, 580
Hoffmann, Katja, 324, 621, 626
Hoffmann, Peter, 616
Hoffmann, Sabine, 184, 235
Hoffmann, Sven, 140
Hoffmann, Verena S, 133, 917
Hoffmann, Wolfgang, 637, 638, 758, 760, 761, 771
Hog, Johannes, 1070, 1079
Hogger, Louisa Marie, 755
Holdgrün, Ulrike, 925
Holsmoelle, Arvid, 728
Holst, Klaus Kähler, 269, 386
Holzlehner, Julia, 922, 923
Hoogland, Jeroen, 338, 534
Hooker, Andrew C., 183, 529
Hooper, Richard, 411, 961
Hoppe, Dorothée B., 975
Horsch, Cornelia, 680
Horstmeier, Lukas M., 882
Hossain, Md. Monowar, 865
Hossain, Syed Shahadat, 402
Hössjer, Ola, 508
Hothorn, Ludwig, 235
Hothorn, Torsten, 531
Hoven, Emiel Van den, 705
Hoyer, Annika, 280, 281
Hu, Lin, 532
Hu, Taojun, 423
Hu, Zheng Jing Jimmy, 1026
Huan, Cheng, 82
Huang, Chenglong, 1075
Huang, Yi, 430
Hübel, Kerstin, 928
Huber, Markus, 456
Hubin, Aliaksandr, 234
Hübner, Dennis, 396, 751
Hübner, Ursula Hertha, 323, 327
Huebner, Marianne, 741, 915
Huei, Sin-Yin, 272
Hügel, Jonas, 440, 983
Hughes, Arthur, 569
Hughes, Rachael, 400
Huirne, Judith, 434
Humphreys, Keith, 853
Hund, Hauke, 629
Hungele, Andreas, 624
Hunt, Alexandra, 967
Hüsing, Paul, 223
Hutfluss, Roman, 765
Hutter, Frank, 493, 1044, 1051, 1070, 1079, 1082
Huys, Isabelle, 820
Iancu, Alexa, 763
Iapaolo, Valeria, 446
Ibrahim, Nurain, 655
Ickstadt, Katja, 184, 185
Idema, Demy, 735
Ieva, Francesca, 253, 294, 337, 344, 446, 556, 581, 692
Igl, Wilmar, 203, 754, 1021
Ignatenko, Yevgeniia, 764, 765
Ignatiadis, Nikos, 426
Ihl, Vanessa, 360
Ihorst, Gabriele, 398
Ikelaar, Nadine A., 707
Illigens, Ben, 689
Im, Jiwon, 901
Immer, Franz, 928
Imtiaz, Sara, 972
Innocenti, Francesco, 546
IntHout, Joanna, 813
Irie, Takumi, 795
Irkinaitė, Agnė, 243
Ishikawa, Yuya, 387
Ishizaka, Shuzo, 696, 949
Ishizuka, Naoki, 1020
Iversen, Aske Thorn, 269
Izmirli, Burakcan, 169
Izumi, Shoki, 428

Jabaudon, Matthieu, 549, 850
Jaberansary, Mehrshad, 100, 101, 102, 103
Jackson, Christopher, 716
Jackson, Dan, 787
Jacobs, Tijn, 479
Jacobsen, Anne, 922
Jacqmin-Gadda, Hélène, 358, 591
Jafari, Hassan, 198
Jäger, Christian, 244, 346, 908, 909
Jäger, Thomas, 808
Jahn, Beate, 593, 699, 1023
Jahn, Franziska, 541, 622
Jain, Arpit, 140
Jaki, Thomas, 94, 95, 181, 451
Jalali, Amirhossein, 1007
Janetzko, Patrick, 247
Jang, Insu, 890
Jani, Ilesh V., 932
Janiaud, Perrine, 955
Jannot, Anne-Sophie, 276, 289, 437, 712, 742, 1003
Jansen, Eva, 931
Jansen, Gunther, 887
Janus, Annette, 960
Japar, Izzati Izyani, 463
Jaskulski, Stefanie, 977
Jay, Nicolas, 276
Jelizarow, Monika, 172, 173, 193
Jenkins, David, 394, 452, 659, 683, 810
Jennison, Christopher, 474
Jennum, Poul Jørgen, 440
Jentsch, Ella Maria, 983
Jepsen, Peter, 974
Jerome, Lambert, 829
Jespersen, Louise, 529
Jesse, David, 819
Jiang, Yunxuan, 405
Jírů-Hillmann, Steffi, 223
Jitu, Md. Hasibul Islam, 402
Jogho, George, 920
Johnson, Autumn, 211
Jones, Hayley, 835
Jörß, Katharina, 540
Joura, Elmar, 593
Jreich, Rana, 388, 482
Jui, Moss. Maharun Nahar, 991
Jullion, Astrid, 866
Jullum, Martin, 567
Jung, Hyun Jun, 493
Jung, Ian-C., 912
Jung, Jaegyun, 270
Jung, Klaus, 144
Jung, Yunchae, 890
Kaase, Martin, 586
Kaat, Aaron, 985
Kabalimu, Anna, 882
Kaboth, Pauline, 398
Kabus, Fabian, 349, 1052, 1060, 1061, 1071
Kacha, Ghanshyam, 844
Kacprowski, Tim, 541
Kadioglu, Dennis, 537
Kadra, Arlind, 1083
Kailer, Jakob, 495
Kaisaridi, Sofia, 706
Kaiser, Vanessa, 882
Kakkassery, Vinodh, 626
Kaletta, Sebastian, 612
Kalweit, Gabriel, 1042
Kalweit, Maria, 1042
Kamal, Vineet Kumar, 693, 798, 934
Kamarudin, Adina, 655
Kamel, Rafiq, 1083
Kammer, Michael, 228, 234, 412, 726
Kammgne, Sorelle Fekam, 191
Kampf, Marvin O., 760
Kamzol, Nicolas A., 630
Kanamori, Takafumi, 815
Kandavel, Thennarasu, 790, 976
Kang, Ji-Hyoun, 405
Kanik, EMİNE, 378, 725, 913
Kapanidis, Evripidis, 853

Kapiteijn, Ellen, 658
Kappe, Miriam, 782
Kappenberg, Franziska, 1098
Kappler, Kerstin, 566
Kapsner, Lorenz, 763
Karakaya, Jale, 588
Karakioulaki, Meropi, 257, 704, 1048, 1068, 1076, 1078
Karbach, Ute, 907
Karimi, Zahra, 844
Karimian, Golnar, 714
Kasap, Zeliha Aydin, 740
Kasbohm, Elisa, 397, 442
Kasenda, Benjamin, 955
Kaspar, Mathias, 802
Kasza, Jessica, 559, 560
Katalinic, Alexander, 1005
Katina, Stanislav, 911
Katsahian, Sandrine, 254, 482
Katsoulis, Michail, 855
Katsumata, Hiroki, 671
Katzensteiner, Matthias, 441, 604, 615
Katzke, Verena, 977
Kaufmann, Masako, 1064
Kaufmes, Kevin, 783
Kauk, Luise, 779
Kauth, Verena, 350
Kawaguchi, Atsushi, 295
Kazaz, Tanyeli Guneyligil, 335
Keane, Benjamin, 743
Kedida, Jiregna Olani, 1008
Kehl, Sven, 971
Keil, Thomas, 803
Keller, Felix, 982
Keller, Martin B., 1084
Kemmerer, Pascal, 685
Keogh, Ruth, 76, 262, 436, 438, 582, 833
Kepenek, Hamdi Furkan, 884, 886
Kern, Jori, 605
Kern-Isberner, Gabriele, 664
Kers, Jesper, 1000
Kesselmeier, Miriam, 753
Kestler, Hans Armin, 606
Kesztyues, Tibor, 541
Ketabi, Ariane, 606, 608
Khalifa, Ottavio, 353
Khalil, Ahmed, 943
Khalil, Hanan, 411
Khan, Md Hasinur Rahaman, 466, 510, 865, 949, 991
Khan, Sam, 869
Khawaga, Ghada El, 663
Khera, Monil, 554
Khunti, Kamlesh, 844, 989
Kiefer, Lena S., 891
Kiel, Alexander, 396, 536
Kieser, Meinhard, 286, 523, 650, 675, 678, 728, 846
Kikuchi, Tsutomu, 696
Kim, Dokyoon, 713
Kim, Jane J., 593
Kim, Sun Bin, 252
Kinast, Benjamin, 112
Kindler, Thomas, 245
Kirschner, Janbernd, 255, 1063
Kirsten, Toralf, 100, 101, 102, 103, 324, 345, 925
Kirton, Laura, 288, 801
Kjaer, Magnus, 962
Klaaßen, Tammo, 778
Kleesiek, Jens, 897
Klein, Christine, 548
Klein, Daniel N., 1084
Kleinert, Thomas, 396
Klimek, Matthias, 993
Klingmueller, Florian, 181, 183, 529
Klopfenstein, Sophie, 114, 116, 127
Klug, Stefanie J., 404
Kluttig, Alexander, 999
Knaup, Petra, 612
Kneuer, Carsten, 874
Knevel, Rachel, 551
Kobak, Eva, 626
Köber, Göran, 407

Kober, Nils, 1046, 1059, 1061
Koc, Gonul Hazal, 405
Koch, Armin, 259, 1001
Koch, Lennart, 593
Koch, Philipp, 299
Koch, Raphael, 791
Kochanowicz, Marta, 1018
Köchert, Karl, 372, 959
Koenen, Niklas, 513
Koenig, Franz, 94, 181, 183, 529, 562, 711
Koglin, Alexander, 637
Kohlbacher, Oliver, 100, 101, 102, 103, 626
Koinig, Karin, 975
Kok, Marc R., 405
Köksal, Cise Nil, 764
Köksal, Ryan, 1047, 1065
Kölbel, Vivian, 398
Koletzko, Lukas, 596
Koliopanos, Georgios, 633
Koller, Michael, 928
Konietschke, Frank, 283
König, Inke R., 548, 973
Konigorski, Stefan, 217
Konstantinidis, Menelaos, 330
Kontos, Dimitris, 673
Kooiman, Judith, 343
Koos, Björn, 566
Kopar, Serli, 708
Kopp, Kevin, 897
Kopp, Nina, 923
Kopp-Schneider, Annette, 67
Koppelin, Frauke, 604
Köpper, Hannah, 914
Kordtomeikel, Kian, 1052, 1060
Kornyushyn, Vasyl, 377
Koshy, Priyanka, 1000
Kossen, Robert, 443, 606, 608
Köster, Helene, 753, 780
Kostouraki, Andriana, 457
Kotter, Elmar, 1069, 1077
Köttgen, Anna, 493, 571, 794, 1044, 1075
Köttgen, Michael, 493
Koumantakis, Emanuele, 940
Kouton, Ludivine, 1003
Koutsiouroumpa, Ourania, 736
Kouyos, Roger, 839
Kramarzyk, Filip, 244
Kramer, Frank, 298
Krampe, Jonas, 275
Kranzusch, Joe, 445
Krauß, Anna, 761
Krawczak, Michael, 803
Kraywinkel, Klaus, 979
Krefting, Dagmar, 171, 194, 443, 540, 604, 608, 745, 757, 918, 942
Krenzien, Felix, 894
Kreuer, Sascha, 808
Kreuzer, Michaela, 980
Kreuzthaler, Markus, 197
Krichevsky, Alina, 759
Krißmer, Sonia Maria, 494, 1047, 1065
Krist, Lilian, 803
Kroidl, Arne, 932
Kröner, Saskia, 112
Krotka, Pavla, 429
Krügel, Nathalie, 928
Kruse, Henner M., 613
Kuchin, Igor, 699, 1023
Kücking, Florian, 323, 327
Kuehni, Claudia E., 609
Kufner, Anna, 943
Kuhn, Christian, 928
Kühn, Nikolas, 190
Kuhn, Peter, 1027
Kühnapfel, Andreas, 864
Kul, Seval, 335, 740
Kulaga, Anton, 145
Kulagin, Valentin, 487
Kumar, Aditya, 225, 355
Kumar, Anup, 1047, 1065
Kundu, Monica, 989

Kuntz, Alessandra Simone, 230, 326
Kunz, Cornelia Ursula, 953
Kurian-Chandy, Delight, 191
Kuroda, So, 816
Kurscheidt, Maximilian, 536, 629
Kurzmann, Melanie, 1057
Kusch, Harald, 168, 632, 778
Kuß, Oliver, 367
Kutafina, Ekaterina, 222, 325
Kuypers, Dirk, 1000
Kwarteng, Nana-adjoa, 331, 1080
L'Imperio, Vincenzo, 497
La Bella, Saverio, 641
La Faige, Marc Gros, 712
la Rubia, Eva Coll-de, 714
Laabs, Björn-Hergen, 548
Laber, Eric, 87
Lacruz, Maria Elena, 999
Lahroudi, Zahra Seyedi, 291, 720
Laible, Marie-Christine, 770
Laight, Jak, 801
Laliena, Antonio R., 809
Lambert, Paul, 339, 467
Lambert, Philippe, 504
Lamoš, Martin, 911
Lando, Stefania, 381, 519, 788, 842, 860, 1008
Landoulsi, Zied, 548
Lang, Benjamin, 953
Lange, Christoph, 797
Lange, Maximin, 198
Lange, Zoe, 596, 1051, 1082
Lange, Zoe Kristin, 555, 1058
Lantz, Anna, 590
Lanwer, Leonas, 997, 1006
Lanzinger, Stefanie, 624
Lapidus, Nathanaël, 435
Lara, Elena Garcia, 813
Large, Steve, 436
LaRosée, Felicitas, 398
Lasinio, Giovanna Jona, 694
Latif, Mahbub A. H. M., 688
Laudien, Martin, 803
Lauer, Romy, 421
Lawes, Mario, 689
Lawless, Caroline, 488
Le Cessie, Saskia, 551, 1093, 1094, 1099
Le Clech, Renée, 276, 712
Le Douce, Juliette, 960
Le Ralle, Svetlana, 698
Le Teuff, Gwénaël, 863
Leddig, Torsten, 760, 761, 771
Ledoux, Johanna, 825
Lee, Hyung-Chul, 270
Lee, J. Jack, 77
Lee, Katherine, 75, 152, 153, 677, 824, 1094
Leegwater, Hanneke, 114, 127
Lees, Charlie W., 721
Legha, Amardeep, 642, 643
Lehmann, Kevin, 1028
Lei, Rayleigh, 741
Lemster, Simon, 235
Lenz, Philipp, 580
Lenz, Stefan, 195, 245, 350
Lesosky, Maia, 827
Letzelter, Philomène, 254
Leufkens, Daniel, 1016
Leurent, Baptiste, 268, 961, 966
Leutenegger, Anne-Louise, 712
Leverkus, Friedhelm, 163
Levin, Leonard, 846
Levy, Sacha, 388
Lewin, Alex, 240, 431
Leyrat, Clemence, 212, 961
Li, Enyu, 362
Li, Huijie, 570
Li, Jialiang, 963
Li, Letao, 453
Li, Liming, 958
Li, Yi, 81
Li, Yong, 493

Li, Zhe, 251
Lichtner, Gregor, 409
Lieb, Wolfgang, 803
Liebe, Jan-David, 323
Liegel, Patrick, 346, 908, 909
Liehr, Darian, 441, 604
Liesche-Starnecker, Friederike, 589
Lilienthal, Jona, 383
Lin, Tzu-Chuan, 669
Lind, Charlotte, 298
Linde, Klaus, 280
Lindemann, Yvonne, 905, 930
Lindson, Nicola, 733
Linger, Anna, 302, 652
Lingjærde, Christoffer, 266
Lingsma, Hester, 292
Lip, Gregory, 835
Lipprandt, Myriam, 626
Litt, Jill S., 699, 837, 1023
Litwin, Tim, 704, 1048, 1076, 1078
Litzcke, Vivian, 923
Liu, Xiaofei, 259
Liu, Yuan, 480
Liu, Yumeng, 680
Liubisheva, Lada, 625
Löbe, Matthias, 114, 117, 127, 168, 230, 396
Lobentanzer, Sebastian, 144
Lodde, Georg C., 792
Löffler, Markus, 375, 753, 780, 899
Logan, Katherine, 581
Löhberg, Christian R, 971
Lohmann, Katja, 548
Löhnhardt, Benjamin, 142, 156
Lohr, Christina, 375, 905
Long, Yongxi, 426
Lopez-Perez, Laura, 640
Lotz, Anne, 854
Lounaci, Kahina, 388
Lozano-Kuehne, Jingky, 743
Lübbert, Michael, 836
Lucas, Tim, 950
Luetz, Alawi, 806
Luime, Jolanda J., 405
Lukic, Ivan, 355
Lumbers, R T, 582
Lunn, Dave, 800
Lüpkes, Christian, 1028
Lusa, Lara, 1093, 1096
Luschei, Remi, 309
Lüser, A. Altar, 894
Lutz, Sharon, 797
Lutz-Bonengel, Sabine, 1051, 1082
Lwilla, Anange F., 932
Ly, Alexander, 426
Lyons, Jane, 554, 996
Lyons, Ronan A, 996
Ma, Chengshan, 367
Ma, Jie, 877
Ma, Wenjuan, 741
MacGregor, Louise, 198
Maetzler, Corina, 803
Maetzler, Walter, 708, 803
Magini, Giulia, 655
Magirr, Dominic, 362, 958
Magnus, Celine, 303
Mahar, Robert, 153, 677
Maharani, Asri, 720
Maher, Francesca, 198
Mahler, Andreas, 765
Mainetti, Marta, 450, 503
Mainzer, Rheanna, 1094
Maiuri, Tomas, 573
Majeed, Raphael W., 247, 605, 904
Malafosse, Hugo, 437
Malakowsky, Jolina, 922
Malats, Núria, 492, 647
Malhotra, Devon, 871
Malik, Mikail, 871
Malkawi, Clemens, 324
Mander, Adrian P., 521

Mang, Jonathan M., 485, 762, 763
Manjunatha, Kaushik, 538, 630
Mansmann, Ulrich, 118, 129, 649, 1105
Manuilova, Iryna, 799
Manzke, Mathias, 608
Mappes, Andreas, 924
Marchioro, Linda, 88
Marcus-Alic, Katrin, 566
Marie, Sébastien, 960
Markham, Sarah, 227, 287, 956
Marschner, Ian, 858
Marschollek, Michael, 189, 634
Marshall, Dominic C, 549
Marson, Anthony, 835
Martens, Eimo, 147, 159, 160, 176, 178
Martens, Robin, 875
Martin, Glen, 642, 835
Martin, Johanna, 925
Martín-Peña, Rosa E., 134, 136
Martinelli, Elena, 640
Martinez, Jose Manuel Lema, 405, 681
Martini, Alberto, 641
Martyn, Meredith, 198
März, Winfried, 759
Marzo-Ortega, Helena, 405
Masci, Chiara, 337
Masó-Aguado, Montse, 699, 837, 1023
Mass, Marco, 299
Massa, Davide, 381
Massa, Sofia, 405
Masseroli, Marco, 446
Masset, Christophe, 1000
Mathé, Philipp, 784
Mathes, Georg, 248, 376, 538, 631
Mathes, Tim, 133, 584
Matsui, Kota, 671, 672, 674, 747, 815
Matsui, Shigeyuki, 662, 671, 672, 674, 747, 815, 957
Matsuyama, Yutaka, 428, 460
Mattio, Francesco Maria, 883
Matus, Dominika, 1086
Maun, Andy, 882, 920
Maurel, Benjamin, 516
Maurer, Miriam Cindy, 745
Maurer, Theresa, 1051, 1082
Maushagen, Juliane, 891
Mavridis, Dimitrios, 734, 736
May, Patrick, 548
Mayer, Benjamin, 680
Mayer, Rudolf, 942
Mazlooman, Seyed Reza, 626
Mbizvo, Gashirai, 835
Mbuagbaw, Lawrence, 561, 1026
McCall, Matthew, 320
McCaw, Zachary, 575
McClenaghan, Elliot, 262
McEvoy, Alannah, 871, 937
McLean, Alistair, 677
McMillan, Brian, 394
McMullen, Justin, 887
Meade, Angela, 314
Meeßen, Corinna, 984
Meggi, Bindiya, 932
Mehlan, Henry, 638
Mei, Yifan, 471
Meineke, Frank A., 375, 769
Meisegeier, Stefan, 1028
Meißner, Yvette, 700
Meiszl, Katharina, 368
Melis, Guadalupe Gómez, 216
Mellac, Katell, 928
Mellmann, Alexander, 586
Melnik, Sima, 275
Melotti, Roberto, 985
Menekseoglu, Pelin Özkara, 907
Meng, Zhaoling, 388
Menger, Jonatan, 494, 1047, 1065
Menssen, Max, 874
Mercier, Francois, 711, 819, 866
Merks, Hans, 864
Merks, Johannes Hm, 480

Merle, David, 626
Mertens, Bart J.A., 707
Messina, Paolo, 249, 379, 768, 960
Meßmer, Liane-Marina, 691
Messuti, Maria Eugenia, 319
Metsch, Jacqueline, 199
Metzger, Florian, 708
Meuleman, Saskia Klein, 434
Meyring, Jan, 670
Meziyerh, Soufian, 1000
Michael, Mader, 1021
Michels, Karin B., 977
Michiels, Stefan, 578
Middleton, Melissa, 153, 677, 824
Miebs, Jasko, 1104
Mieno, Makiko, 859
Mikayelyan, Astghik, 754
Mikkelsen, Hjalte Søberg, 465
Mikolajczak, Jakub, 749
Milano, Francesca, 1019
Milazzo, Bianca Maria, 1019
Mileika, Justas, 243
Miljukov, Olga, 803
Milmantas, Paulius, 243
Minusa, Shunsuke, 796, 949
Mioskowski, Miguel, 350
Miranda-Afonso, Pedro, 343
Mirraziroudsari, Seyedeh Delaram, 626, 773
Mischinger, Lukas, 282
Mishian, Hamed, 621
Mitra, Prasenjit, 124, 125
Mittnacht, Lukas, 761
Moazemi, Sobhan, 134, 136
Modat, Marc, 1009
Modersohn, Luise, 196, 347, 375
Moffa, Giusi, 840, 955
Moghaddam, Shirin, 1007
Mohammed, Mohanad Mohammed Adam, 933
Moher, David, 722
Moissl-Blanke, Angela P., 759
Mokhtari, Nour Eddine El, 803
Moler-Zapata, Silvia, 489
Molina, Kevin Alejandro, 443
Möllenhoff, Kathrin, 406
Möller, Aleyna, 923
Möller, Sebastian, 894
Monasta, Lorenzo, 660
Monchietto, Giulia, 946
Monnereau, Matthias, 578
Moons, Karel, 391, 419, 602, 642, 735
Moradpour, Maryam, 514
Morbach, Caroline, 803
Moreno-Betancur, Margarita, 75, 433
Morgan, Katy, 833
Mori, Keiko, 727
Mori, Tim, 133
Moritz, Regina, 618
Moriuchi, Hiroaki, 830
Morris, Tim, 361, 412, 575
Moser, Frank-Michael, 772
Mossop, Helen, 521
Moter, Alina, 1042
Mougeni, Fabrice Lotola, 310
Mozgunov, Pavel, 313, 524
Mozun, Rebeca, 609
Mühlberger, Nikolai, 593
Mühler, Milena, 593
Muilwijk, Danya, 471, 820
Mukowski-Kickhöfel, Rebecca, 223
Mulder, Michelle L.M., 405
Müller, Dominik, 298
Müller, Julia, 1080, 1084
Müller, Lutz P., 541
Müller, Marcel, 118, 129
Müller, Marilena, 478
Müller, Mattea, 189
Müller, Thomas, 928
Müller, Tim, 926
Müller, Wiebke, 823
Müller, Wolfgang, 168, 169

Müller-Wirtz, Lukas, 808
Munch, Anders, 574
Munn, Zachary, 411
Murphy, John D., 724
Murray, James, 436
Murunga, Nickson, 467
Musisi, Connie, 888
Muzoora, Michael Rusongoza, 326
Mychajliw, Christian, 708
Naesens, Maarten, 1000
Nagai, Luis, 144, 145
Naji, Leen N, 871
Najork, Solveig, 299
Nakagawa, Tohru, 796
Nakamura, Yusuke, 815
Nanjaiah, Nandakumar Dalavaikodihalli, 976
Naor, Elisheva, 191
Narendran, Parth, 1019
Naue, Jana, 1051, 1082
Naumann, Laura, 323
Navarro, Jùlia Santamaria, 699, 1023
Nazzari, Valeria, 581
Ndonfack, Felix Barrez Tambe, 1062, 1063
Neema, Faith, 876
Nematollahi, Iman, 1042
Nensa, Felix, 538, 630, 897
Neschkova, Sneschana, 941
Nesteruk, Marta, 497
Nettekoven, Charlotte, 398
Neukum, Manuel, 274
Neumann, Daniel, 484, 611, 753, 780, 899
Neumayer, Hannah, 217
Neven, Anouk, 181
Nevill, Clareece, 733
Ngueda, Christelle Kemda, 326
Nguyen, Thuong-Thuong, 1009
Niazmand, Emetis, 541
Nicole, Eter, 626
Nieboer, Daan, 292
Niederseer, David, 636
Niederwieser, Dietger, 611
Niehoff, Nicole, 580
Nielsen, Kari, 978
Nieto-Gutierrez, Wendy, 489
Nikolakopoulos, Stavros, 600, 736
Nikolakopoulou, Adriani, 331, 332, 734, 1080
Nikolic, Miki, 440
Niks, Erik H., 707
Nishi, Suzuna, 747
Nist, Anna M., 1014
Noguchi, Yoshihiro, 684
Noirjean, Silvia, 500
Nold, Mariana, 234
Nolde, Janis, 417
Nomura, Shogo, 428
Nonnenmacher, Tobias, 1017
Norajitra, Tobias, 1017
Nordland, Andreas, 269
Nowak, Hartmuth, 147, 159, 176, 178, 179, 566
Nowowiejska-Wiewióra, Alicja, 777
Nsengimana, Jérémie, 743
Nuhic, Sara, 1036
Nürnberger, Carolin, 803
Nussbaumer, Maximilian, 621
Nussbeck, Sara Yasemin, 632
Nüßler, Frank, 626
Nuxoll, Lisa-Marie, 973
Nyoungui, Elisabeth Félicité, 942
Nyström, Per, 962
O'Connor, Daniel, 1031
O'Keeffe, Aidan, 974
Oba, Koji, 460
Obi, Sean, 227
Ocampo, Alex, 459, 575
Oefner, Peter J., 794
Oehm, Johannes, 326, 586, 626
Oehring, Robert, 894
Oeltze-Jafra, Steffen, 541
Oest, Lars, 624
Oesterle, Sabine, 73, 114, 115, 127

Ogawa, Mitsunori, 428
Ohigashi, Tomohiro, 815
Okamura, Satomi, 818, 1025
Oksanen, Anna, 978
Olaniran, Oyebayo Ridwan, 227, 287, 956
Oliveira, Cláudia Cruz, 292
Olsen, Inge Christoffer, 94, 95, 96
Olszanecka-Glinianowicz, Magdalena, 1018
Omar, Rumana Z, 463
Omeirat, Jazia, 538, 630
Omori, Takashi, 679
Ong, Sean, 153, 824
Oppenberg, Jan, 770
Oppong, Felix, 202
Oprea, Anna, 871
Oremek, Maximilian, 194
Orenti, Annalisa, 438
Orlovsky, Henry, 806
Ortholand, Juliette, 576
Ortmann, Natalia, 764, 765
Osaki, Takanobu, 696, 949
Osborne, Timothy, 373, 554
Osman, Farzana, 933
Osmanodja, Bilgin, 225
Ott, Manuela, 263, 499, 667
Otto, Christian, 623
Otto, Kay-Michael, 168, 170
Otto, Lisa Rebecca, 223
Overton, Philip, 396
Owczarek, Aleksander, 1018
Owen, Rhiannon K, 354, 373, 554, 715, 996
Oyewale, Michael, 786
Ozkan, Asuman Deveci, 884
Ozturk, Ebru, 884
Özültürk, Pınar, 935
Pach, Daniel, 217
Pacynko, Samantha, 452
Paganoni, Anna, 337
Pagkalidou, Eirini, 330
Pagliazzi, Angelica, 1000
Pal, Anita, 934
Pal, Pratyasha, 1031
Pallaoro, Peter, 347
Palm, Julia, 930
Pamminger, Moritz, 228, 726
Panagiotopoulou, Kanella, 334
Panagiotou, Gianni, 977
Pandit, Harshvardhan J., 230
Panesar, Balpreet, 330
Panholzer, Torsten, 195, 245, 350
Pap, Ákos Ferenc, 511
Papakonstantinou, Eleni, 417
Papanikos, Tasos, 393
Pape, Mona, 604
Papenberg, Martin, 668
Papon, Veronika, 699, 1023
Paprott, Rebecca, 217
Parast, Layla, 569
Pargeter, William Jobson, 1064
Park, Minsu, 807, 890, 901
Parmar, Mahesh Kb, 314, 680
Parpia, Sameer, 871
Parr, Harry, 393, 852
Parsons, Marc, 561
Partington, Giles, 737
Pattaro, Cristian, 415, 985
Pau, David, 437, 698
Paulet, Lucian, 626
Pautonnier, Joanna, 861
Pavillet, Clara, 131
Pavlou, Menelaos, 463, 645
Pawłowski, Marco, 744
Pawel, Samuel, 731
Pearson, Oliver, 198
Pečeliūnas, Valdas, 243
Pechmann, Astrid, 255, 1063
Pelka, Obioma, 538, 630
Pelz, Anne, 273
Penzel, Thomas, 918
Perego, Gaia, 1019

Pereira, Bruno, 850
Perez, Ignacio Gonzalez, 547
Perlino, Federico Luigi, 365
Perme, Maja Pohar, 507, 592
Pernickel, Sophie, 191
Perperoglou, Aris, 393, 800, 852, 1091
Perry, Benjamin, 91
Perz, Leander, 941
Pessia, Alberto, 673
Pétavy, Frank, 181
Peter, Kimani, 362
Peters, Annette, 686, 891
Petit, François, 353, 573
Petrasch, Gloria, 636
Petrof, Oana, 800
Petroff, David, 864, 987
Pfaffelhuber, Peter, 352, 1051, 1058, 1081, 1082
Pfaffenlehner, Michelle, 124, 125, 1061, 1063
Pfefferle, Alexander, 1070, 1079
Philipona, Frederik, 450, 503
Philipps, Viviane, 205
Phillippo, David, 163, 164
Phillips, Daniel J, 988
Pieper, Alexander, 527, 676
Pieper, Dawid, 411
Pigeot, Iris, 789
Pilz, Maximilian, 1013
Piper, Sophie K., 806
Pipper, Christian Bressen, 386
Pirmohamed, Munir, 354
Pirondini, Leah, 833
Placke, Jan-Malte, 792
Placzek, Peter, 100, 101, 102, 103
Pleyer, Jonas, 1049, 1072
Plicka, Philipp, 622
Plieger, Thomas, 916
Plotzky, Christian, 920
Poikela, Maija, 326
Poirot-Bourdain, Antoine, 289
Poiseuil, Marie, 591
Poli, Tito, 640
Pons-Kühnemann, Jörn, 570, 1016
Poole, Charles, 724
Popp, Anita, 274
Posch, Martin, 181, 310, 429, 711, 866
Posfay-Barbe, Klara, 609
Post, Richard, 476
Prague, Mélanie, 251, 501
Pramstaller, Peter, 985
Prasser, Fabian, 100, 101, 102, 103
Presterl, Elisabeth, 941
Pricci, Flavia, 1019
Priest, Lee, 682
Prinz, Katharina, 333
Pritzkuleit, Ron, 1005
Prochaska, Eveline, 614
Procureur, Adrien, 752
Prokosch, Hans-Ulrich, 147, 159, 176, 178
Prosepe, Ilaria, 551, 552
Proust-Lima, Cécile, 205, 251, 265, 358
Prus, Sascha, 626, 628, 773
Pryce, Matthew, 76
Pryss, Rüdiger, 223
Pucher, Gernot, 897
Puljak, Livia, 411
Punjani, Tanvi, 937
Puntscher, Sibylle, 328, 699, 837, 1023
Purohit, Anne-Maria, 907
Purucker, Lennart, 784, 1048, 1051, 1070, 1078, 1079, 1082, 1086
Putter, Hein, 107, 109, 110, 317, 318, 476, 552
Quell, Dorian, 764
Querfurt, Maeve, 922
Quinquis, Laurent, 831
Rabbee, Nusrat, 731
Rabe, Klaus, 803
Rachel, Timo, 704, 1076
Radermacher, Walter J., 184, 185
Radicke, Franziska, 606
Rafferty, James, 354, 373, 554, 715
Raffler, Johannes, 765

Rahman, Shafiqur, 463
Rahmel, Tim, 566
Rahnenführer, Jörg, 406, 792, 1096
Ramasetti, Nikitha Shruthi, 894
Ramjith, Jordache, 320
Rank, Nina, 625
Rapsomaniki, Eleni, 517
Rasheed, Abdur, 972
Rashid, Sheikh Faisal, 134, 137
Rasmussen, Kristine Bjarnum, 974
Rast, Eilin, 770
Räther, Henrike, 222
Ratta, Marco, 313
Ratti, Edoardo, 365
Rau, Moritz, 924
Rauch, Geraldine, 184, 187
Rauh, Johannes, 881
Raum, Hanne, 1071, 1074
Räuscher, Editha, 232
Ravichandran, Ajay Madhavan, 225
Ravn, Henrik, 507
Razieh, Cameron, 844, 939
Rebora, Paola, 710, 992
Recker, Florian, 931
Reese, Jens-Peter, 803
Rehse, Jannik, 281
Reich, Christoph, 691
Reich, Olaf, 593
Reichert, Manfred, 606
Reimhold, Thomas, 931
Reinhardt, Christian, 897
Reith, Wolfgang, 748
Remer, Thomas, 570
Remoundou, Konstantina, 940
Ren, Siyuan, 679
René, Théo, 501
Rentsch, Christopher T., 690
Reusche, Matthias, 614, 780
Revaz, Marys, 840
Rey, Iñaki Soto, 764
Reynders, Dries, 206, 1094
Reynolds, Evan, 415
Rezaei-Darzi, Ehsan, 559
Rezniczek, Günther, 700
Rheinländer, Sophia, 778, 983
Ricci, Felipe, 730
Riccio, Cristian, 633
Richardson, Sylvia, 266, 565
Richter, Adrian, 700
Richter, Jendrik, 540, 604, 757
Richter, Jutta G., 700
Richter-Pechanski, Phillip, 906
Rieg, Siegbert, 784
Rieß, Eva-Maria, 443
Rieß, Jan, 636
Riesterer, Karla, 770
Rigaux, Clemence, 959
Riley, Richard, 288, 391, 419, 462, 599, 602, 642, 643, 801, 869, 877
Riley, Stephanie, 683
Ringwald, Friedemann G., 612
Ringwald, Kai, 834
Ríos, Ana Sofía, 943
Rißner, Florian, 613
Ristl, Robin, 237, 529, 941
Ritchie, Marylyn, 65, 713
Ritter, Kerstin, 708
Ritter, Zully, 134, 514
Ritz, Alexander, 678
Rizopoulos, Dimitris, 202, 343
Rizoudis, Alexia, 298
Rjeschni, Sandra, 932
Rmo, Mustafa, 873
Robert, Marius Vachia, 591
Roberts, Emily, 1090
Robinson, Mark, 142, 156
Robledo, Kristy, 858, 1103
Rocchi, Marco, 556
Rocci, Roberto, 848
Rochau, Ursula, 699, 837, 1023
Röder, Daniel, 147, 159, 161, 176, 178

Rodrigues, Myanca, 330, 871, 937
Rodriguez-Girondo, Mar, 552
Roes, Kit, 172, 193, 813
Roeshammar, Daniel, 960
Rogge, Alizé, 223
Rohde, Angelika, 417, 1046, 1059, 1081
Röhl, Alexander, 100, 101, 102, 103
Röhrig, Rainer, 147, 159, 176, 178, 639
Roig, Marta Bofill, 429
Roll, Stephanie, 931
Roller, Roland, 134, 135, 197, 225, 767, 894
Rollin, Johan, 494
Roman-Ulrich, Müller, 625
Römhild, Jörg, 760
Rondeau, Virginie, 591
Roobol, Monique J., 292
Roos, Malgorzata, 263, 499, 666, 667
Rosa, Ester, 381, 519, 788, 842, 860, 1008
Rosario, Lakisha Ortiz, 195, 245
Rosato, Stefano, 694
Rösch, Alexander, 792
Rose, Matthias, 223
Röseberg, Carl Philip, 445
Rosetta, Lara, 656
Rosinach, Núria Queralt, 114, 116, 127
Rospleszcz, Susanne, 686, 891, 1017
Rostalski, Till, 897
Röttger, Richard, 942
Rouette, Julie, 262
Rouhou, Amina Cheikh, 846
Roussaki, Ioanna, 940
Routen, Ash, 989
Röver, Christian, 133, 866, 874
Rowlands, Alex V., 939
Roy, Pascal, 861
Rubarth, Kerstin, 931
Rubinski, Anna, 759
Rübsamen, Nicole, 544, 1011
Rudi, Tatjana, 700
Rufibach, Kaspar, 261
Rühle, Mathias, 614, 629
Rühlicke, Stefan, 134
Rüland, Thomas, 626
Rüllich, Anne, 905
Rump, Katharina, 566
Ruperto, Nicolino, 641
Ruß, Anne-Kathrin, 803
Rustand, Denis, 358, 719
Rüther, Ferenc Darius, 586
Rutherford, Mark, 339, 467, 845
Ryan, David, 582
Ryan, Oisín, 951
Sabi, Issa, 932
Sabroso-Lasa, Sergio, 492, 647
Saccani, Sebastiano, 1030
Sadeghirad, Behnam, 871
Sadlonova, Monika, 398
Sagara, Issaka, 320
Sahlmann, Jörg, 892
Sai, Nanae, 613, 905
Sakamaki, Kentaro, 815
Salanti, Georgia, 279, 450, 503, 870
Saldeitis, Katja, 632
Saleh, Kutaiba, 613
Salem, Joe-Elie, 752
Salem, Susanna, 584
Salgado, Eduardo, 806
Sallaberry, Marie-Noëlle, 1029
Salm, Claudia, 882
Salogni, Elena, 397, 442
Sam, Jeeva, 247
Samaan, Zainab, 871, 937
Samara, Myrto, 332
Samieri, Cécilia, 251
Sampri, Alexia, 689
Sander, Anja, 149
Sangeux, Morgan, 840
Sarangdhar, Mayur, 684
Saravanan, Ponnusamy, 790
Säring, Dennis, 744

Sartori, Pietro, 636
Sasahara, Yusuke, 830
Saß, Julian, 396, 536
Sato, Tomoharu, 818
Sauer, Christopher, 177, 897
Sauer, Lukas D, 678
Sauerbrei, Willi, 220, 1091, 1096, 1100
Sautier, Caroline, 224
Savelieva, Marina, 453
Savirón-Cornudella, Ricardo, 809
Sawada, Masashi, 662
Sax, Ulrich, 142, 156, 272, 632, 778, 983
Saxer, Franziska, 458
Sayegh-Jodehl, Sabine, 781
Sboui, Moenes, 703
Scarale, MariaGiovanna, 1019
Schaaf, Thomas, 566
Schaarschmidt, Frank, 585, 874
Schaarschmidt, Marco, 326
Schächter, Clemens, 255, 705, 1061, 1062, 1063
Schack, Marcel, 277
Schadendorf, Dirk, 792
Schaeffner, Elke, 794
Schäfer, Anna, 803
Schäfer, Antonia, 655
Schäfer, Henning, 724
Schäfer, Lisa, 925
Schaffrath, Judith, 541
Schandelmaier, Stefan, 411, 722
Schaub, Stefan, 928
Schauberger, Gunther, 404
Schaumburg, Frieder, 791
Scheel, Heiko, 443
Scheike, Thomas, 509, 598
Schemet, Lena, 238
Scherag, André, 613, 753, 899, 930
Schernhammer, Eva, 593
Schewe, Christiane, 147, 148, 159, 176, 178
Schicktanz, Silke, 918
Schiefenhövel, Fridtjof, 231
Schiller, Emily, 765
Schlapbach, Luregn J., 609
Schlattmann, Peter, 104
Schlett, Christopher, 891, 1017
Schlieker, Laura, 372
Schliemann, Antje, 1005
Schlosser, Pascal, 493, 1044, 1070
Schlüter, Miriam, 223
Schmidbauer, Lena, 803
Schmidt, Carsten Oliver, 235, 397, 442, 751, 1093
Schmidt, Daniel, 749
Schmidt, Florian, 484, 753, 780
Schmidt, Marvin, 537
Schmidt, Maximilian, 763
Schmidt, Rene, 817
Schmidt, Thorsten, 1062, 1063
Schmidts, Miriam, 1057
Schmitt, Sabrina, 522
Schmitz, Julika, 1051, 1082
Schmoor, Claudia, 836
Schmutz, Remo, 964
Schneider, Antonius, 280
Schneider, Carolin, 984
Schneider, Matthias, 700
Schneiderheinze, Hauke, 760
Schnepf, Betina, 898
Schöbel, Christoph, 918
Schobel, Johannes, 782, 1027
Schoberth, Mona, 493
Schofield, Georgie, 835
Scholz-Kreisel, Peter, 980
Schomburg, Dirk, 283
Schönfeld, Norman, 616
Schork, Karin, 566
Schorr, Christian, 748, 808
Schrader, Andreas, 919, 922, 923
Schramm, Elisabeth, 1084
Schramm, Sara, 368
Schreiber, Stefan, 803
Schreiweis, Björn, 112

Schröder, Daniel, 626, 773
Schroeder, Jan, 625
Schubert, Ronny, 769
Schuhknecht, Felix, 664
Schuit, Ewoud, 602
Schuldt, Ronny, 760, 761
Schuler, Franziska, 586
Schuler, Katharina, 324, 912
Schulmeister, Sophia, 190
Schult, Thomas, 760
Schulte-Althoff, Matthias, 409
Schultheiss, Ulla T, 571, 794
Schultz, Bruce, 100, 101, 102, 103
Schultze, Anna, 690
Schulz, Maxi, 529, 584
Schulz, Nadine, 604
Schulz, Stefan, 197
Schulze, Jürgen, 177
Schulze, Susann, 799
Schumann, Laura, 1005
Schunda, Patrick, 924
Schure, Judith ter, 427, 724
Schürmeyer, Leonie, 731
Schuster, Alexander, 626
Schuster, Verena, 942
Schwab, Simon, 655, 928
Schwane, Vera, 191
Schwarting, Andreas, 664
Schwarz, Julian, 907
Schwarzer, Guido, 331, 1080
Schwarzwäller, Marius, 700
Schweizer, Simon Tobias, 538, 629
Schwenke, Johannes, 94, 97, 955
Schwinn, Johanna, 802
Scott, Darren, 240, 431
Seaman, Shaun, 448
Sebastien, Bernard, 482
Sedki, Mohammed, 1091
Sedlmayr, Brita, 912
Sedlmayr, Martin, 324, 621, 912
Seeber, Gesine, 639
Seeger, Werner, 247
Seeha, Suteera, 347
Seidling, Hanna M., 779
Seitidis, Georgios, 600, 736
Sekula, Peggy, 794, 977, 1097
Selbmann, Anne Christine, 398
Selles, Ruud, 786
Selman, Chris, 152, 153, 677, 824
Sendel, Sebastian, 548
Senn, Stephen John, 312
Sergeant, Jamie, 394, 851
Serruys, Patrick W., 341
Sethi, Ruben, 372
Seyderhelm, Friederike, 757
sh, maryam, 888
Shah, Anoop, 582
Shamsutdinova, Diana, 227, 252, 287, 956
Shapland, Chin Yang, 400
Sharma, Kuldeep Kumar, 790
Sharma, Sraddha, 322
Sharples, Linda, 436
Sheikhalishahi, Seyedmostafa, 802
Shekhar, Prem Kant, 832
Shibata, Taro, 1020
Shiely, Frances, 722
Shin, Woosub, 820
Shinozaki, Tomohiro, 460
Shoji, Masatoshi, 209
Siddiqui, Masood, 322
Sidler, Daniel, 928
Siebert, Uwe, 328, 593, 699, 837, 1023
Sigle, August, 352
Sigle, Stefan, 248, 376, 538, 631, 783, 1036
Signorelli, Mirko, 340
Sihombing, Giovan Christoffel, 944
Simon, Philipp, 802
Simpson, Julie, 677
Sinclair, Karen, 946
Singh, Janharpreet, 869

Singh, Nimesh, 538, 630
Singh, Tarini, 453
Sirna, Federico, 205
Sitek, Barbara, 566
Skipka, Guido, 383, 731
Skrzypek, Michał, 777
Skubala, Adam, 142, 156, 157, 372
Smith, Aiden, 339
Smith, Ellesha, 733
Smolorz, Sabine Andrea, 632
Smyrnakis, Michail, 1031
Snell, Kym, 288, 642, 643, 644, 801
Sobe, Kathrin, 614
Solano, Pamela, 451
Solmi, Francesca, 500
Soloveva, Polina, 606, 608
Solovieff, Nadia, 887
Solveen, Henrike, 304
Soman, Arya C, 1007
Sommer, Julia, 931
Song, Xinyuan, 82
Sonuvar, Esra Tokur, 345
Soo, Ross Andrew, 963
Sood, Meemansa, 453
Soon, Yu Yang, 210
Soranzo, Nicole, 253
Sordo, Anna, 842, 1008
Soto-Rey, Iñaki, 765
Spain, Thomas, 849
Späthling-Mestekemper, Susanna, 700
Sperrin, Matthew, 452, 659, 683, 835
Spicher, Nicolai, 745
Spies, Claudia, 806
Spieß, Ellena, 589
Spitali, Pietro, 707
Splittgerber, Tom, 513
Sprang, Maximillan, 134, 138
Spreafico, Marta, 480, 656, 864
Spreckelsen, Cord, 623
Spreuer, Til Karsten, 247, 904
Sreekanth, Jayanth, 806
Sriram, Sameeksha, 272
Sroczynski, Gaby, 328, 593, 699, 1023
Stahl, Andreas, 626
Stahl, Dana, 760, 761, 771, 772
Stahl, Daniel, 227, 252, 287, 956
Stahlmann, Katharina, 545
Stahmann, Alexander, 585
Stallard, Nigel, 362, 711, 866
Stamate, Daniel, 252
Stamoulopoulos, Achilleas, 515
Stang, Andreas, 685, 724
Stanislas, Virginie, 977
Stark, Anna, 897
Stark, Didem, 708
Starke, Paula, 529, 584
Stäubert, Sebastian, 753, 780, 899
Stein, Alexandra, 761
Stein, Thomas, 1077
Steinbrenner, Inga, 794
Steinbrinker, Tabea, 540
Steiner, Moritz, 770
Steinhaus, Kristin, 745
Steinmaier, Moritz, 1049, 1072
Steinsaltz, David, 988
Stelzer, Dominikus, 892
Stenzhorn, Holger, 618, 760
Stepanek, Lubomir, 597, 856
Stewart, Derek, 722
Steyerberg, Ewout, 90, 121, 267, 390, 602, 646, 658, 786, 951, 1096
Stiggelbout, Anne, 292
Stołtny, Dorota, 777
Stock, Christian, 172, 174, 193, 384
Stoffregen, Robert, 487
Stojanov, Thomas, 786
Stolz, Daiana, 349, 417, 1052, 1060, 1066
Stolze, Gabriel, 621
Storck, Michael, 620, 628, 773, 900
Storf, Holger, 273, 537, 774, 775
Storjohann, Alex, 922, 923

Störk, Stefan, 540, 803
Storp, Jens, 900
Stoye, Ferdinand Valentin, 280, 281
Strand, Fredrik, 590, 853
Strangfeld, Anja, 700
Strantz, Cosima, 799
Strate, Jana, 323
Straumann, Lisa, 928
Stremel, Oliver, 625
Strizek, Brigitte, 931
Strobel, Alexandra, 133
Strohmaier, Susanne, 228, 726
Strojek, Krzysztof, 777
Strotbaum, Veronika, 112
Strübing, Alexander, 484, 753, 780
Struckmann, Stephan, 442, 751
Strüfing, Lucy, 232
Stuart, Elizabeth, 455
Stübs, Gunthard, 637, 638, 758
Stump, Shura Roman, 538
Sturtz, Sibylle, 731
Stürzinger, Ueli, 928
Stüwe, Louisa, 489, 1022
Su, Xinrui, 645
Sucker, Antje, 792
Sud, Yaroslav, 377, 968
Suhre, Karsten, 891
Sulimma, Bente M., 922
Sultana, Nasrin, 510
Sultana, Shirin, 466
Sünkel, Ulrike, 708
Sutton, Alex J, 280, 733
Svensson, David, 962
Sweeting, Michael, 787
Swenne, Annika, 397
Swinkels, Sophie, 218, 831
Syriopoulou, Elisavet, 339, 508, 978
Szabo, Alina, 749
Szurewsky, Chloé, 863
Tabernig, Barbara, 931
Tackney, Mia, 84
Taeger, Dirk, 854
Taggart, Joseph, 715
Tai, Bee Choo, 210, 963
Tai, Ka Hin, 118, 129
Taimr, Pavel, 1002
Takahashi, Kanae, 1013
Takahashi, Katsuei, 998
Takeda, Kentaro, 260
Takeuchi, Ayano, 727, 795, 990, 998
Tamano, Shu, 369, 387
Tanaka, Noriko, 859
Tanaka, Takeshi, 796
Tang, Yunyou, 943
Tasdogan, Alpaslan, 792
Tassistro, Elena, 994
Tatari, Mahin, 694
Taumberger, Nadja, 593
Tauwaldt, Jan, 190
Tavares, Erick, 1022
Tawiah, Richard, 560
Taylor, Jeremy, 1090
Tazare, John, 262, 690, 833
Tchetverikov, Ilja, 405
Teerenstra, S., 813
Teichmann, Mara, 907
Tellbach, Joshua Georg, 191
Tello, Khodr, 247
Ternès, Nils, 372
Thabane, Lehana, 871, 1026
Thalmann, Paul, 523
Thaqi, Agim, 928
Thas, Olivier, 888
Theen, Sarah, 761
Theillière, Camille, 850
Thevapalan, Andre, 664
Thévenet, Vincent, 1022
Thiebaut, Anne, 1091
Thiebaut, Rodolphe, 488, 569, 828, 1022
Thiel, Sandra, 421

Thies, Saskia, 999
Thirumoorthy, Chinnasamy, 790
Thomas, Nele Henrike, 259, 755, 1001
Thomas, Philippe, 894
Thomassen, Doranne, 121, 1093, 1099
Thompson, Douglas, 393
Thompson, Jennifer A, 961
Thun, Sylvia, 114, 127, 326
Thürwächter, Sina, 783
Thwaites, Guy, 1009
Tibaldi, Fabian, 500
Tibyampansha, Dativa, 664
Tiller, Daniel, 541
Tilling, Kate, 215, 400, 533, 687
Timmer, Antje, 639
Timmerman, Dirk, 646
Timmesfeld, Nina, 368, 421, 695, 832
Tinè, Mariaenrica, 519
Tinsel, Iris, 920
Tippmann, Patric, 420
Tischhauser, Bruno, 928
Tischkowitz, Marc, 661
Tokic, Marianne Charlotte, 421
Tom, Brian, 565
Tomlinson, Laurie, 833
Tomo, Yui, 369
Tony, Roshni, 976
Torchia, Mario, 249, 379, 768, 960
Torri, Vittorio, 344
Totis, Niccolò, 960
Touloumi, Giota, 515
Touré, Vasundra, 114, 115, 127
Tran, Maylis, 576
Tran, Viet-Thi, 353
Travis, James, 172, 193
Tresch, Achim, 406
Trescher, Saskia, 104
Tricco, Andrea C., 330
Triché, Laurent Renard, 549, 850
Trivedi, Srushhti, 134, 139, 918
Troglio, Alina, 625
Truong, Nicolai, 942
Tsegaye, Biruk, 877
Tsokani, Sofia, 330
Tsonaka, Roula, 707
Tsuchida, Jun, 815
Tümler, Kirstin, 624
Turgal, Ebru, 653, 935
Turkiewicz, Aleksandra, 986
Turner, Becky, 599
Tute, Erik, 540
Twerenbold, Raphael, 633
Twrdik, Alexander, 324
Tzortzakis, Stylianos, 857
Uçar, Ali, 935
Ückert, Frank, 608
Ueffing, Marius, 626
Ugurel, Selma, 792
Uiterweer, Emiel Post, 434
Ullmann, Theresa, 1098
Ullrich, Evelyn, 1042
Ulrich, Hannes, 298
Unni, Deepak, 114, 115, 127
Unterberg, Matthias, 566
Urach, Susanne, 183, 529
Ursino, Moreno, 453, 471, 472, 488, 501, 516, 670, 717, 820, 828, 1022
Uschner, Diane, 819
Usher-Smith, Juliet, 661
Ustjanzew, Arsenij, 245, 799
Vach, Werner, 458, 1010
Valada, Abhinav, 1042, 1068
Valberg, Morten, 476
Valenta, Zdeněk, 1002
Valindria, Vanya, 296, 944, 945
Vallejos, Catalina, 721
Valsecchi, Maria Grazia, 365, 994
Valtanen, Mikko, 477
van Amsterdam, Wouter, 951
Van Calster, Ben, 90, 91, 121, 122, 341, 390, 646, 1096
van Dam, Martijn, 391

Van de Velde, Helgi, 388
van de Wiel, Mark, 338, 534, 567
van der Ent, Kors, 471, 820
van der Knaap, Marjo, 502, 811
van der Pas, Stéphanie, 479
van Eekelen, Rik, 434, 502, 811
Van Eijk, Ruben, 600
van Geloven, Nan, 90, 1093
van Goor, Sophia, 864
van Herwaarden, Noortje, 405
van Klaveren, David, 292, 341
Van Oudenhoven, Floor, 218, 577
van Rosmalen, Joost, 427, 577, 735
van Smeden, Maarten, 267, 390, 391, 577, 642, 786, 951
van Straaten, Alexis, 289
van Voorst, Romy, 502
van Wieringen, Wessel, 479
van Zwet, Erik, 426
Vandemeulebroecke, Marc, 711, 866
Vanhoutte, Thomas, 1000
Vannier, Iphigenie, 470
Vansteelandt, Stijn, 76
Vantini, Simone, 692
Varadhan, Ravi, 455
Varghese, Julian, 147, 159, 176, 178, 180
Varoquaux, Gaël, 1022
Varvarà, Giulia, 118, 119, 129
Vasseur, Jessica, 273, 774, 775
Vaulet, Thibaut, 1000
Vayani, Aisha, 845
Vayani, Amber, 950
Vehreschild, Jörg Janne, 803
Velasco-Pardo, Victor, 721
Vellas, Dylan, 1003
Vengadeswaran, Abishaa, 537
Venkatasubramaniam, Ashwini, 416
Venkatesh, Rasika, 713
Vens, Maren, 149, 150
Verbeeck, Johan, 152, 154, 888
Verbeke, Geert, 1000
Vergani, Andrea Mario, 446
Veroniki, Areti-Angeliki, 330
Versloot, Judith, 1026
Vervier, Kevin, 887
Vescovo, Mariavittoria, 1008
Vey, Johannes, 286
Vickers, Andrew, 646
Victor, Anja, 261
Vidal, Maria-Esther, 72
Videm, Pavankumar, 1047, 1065
Viechtbauer, Wolfgang, 867
Viehweger, Elke, 840
Villar, Sofía S, 79, 84, 563
Villena-Ossa, Jose Francisco, 1042
Villoslada-Blanco, Pablo, 492
Vilsmeier, Johannes, 711
Vilsmeier, Judith, 836
Vincentini, Olimpia, 1019
Vis, Marijn, 405
Visconti, Alessia, 940
Vishnevskaya, Valeriya, 248
Vitelli, Valeria, 83, 266
Vittrant, Benjamin, 254
Vogel, Carsten, 223
Vogel, Frederike, 544, 1011
Vogel, Mandy, 606
Vogel, Tanja, 494, 495
Voigt-Radloff, Sebastian, 882
Voller, Corey, 474
Vollmar, Georg, 245
Vollmer, Marcus, 745
von Arnim, Christine A.F., 398
von Bomhard, Nikolaus, 118, 129
von Borzestowski, Marcel, 780
von Felten, Stefanie, 825
von Heydebreck, Anja, 261
von Itter, Marc-Nicolas, 1017
von Schwerin, Reinhold, 589
von Thaler, Anna-Katharina, 708
Vonthein, Reinhard, 969

Voran, Jakob Christoph, 298
Vorisek, Carina Nina, 114, 127, 326
Voß, Sebastian, 372
Vossoughi, Mohammad, 676
Vranken, Arthur, 1000
Vratanar, Bor, 592
Vrijkotte, Tanja, 789
Vs, Binu, 790
Vuorinen, Axel, 472
Vural, Zeynep, 923
Wagenpfeil, Stefan, 449
Wagner, Johnannes, 760
Wagner, Philippe, 805
Waite, Timothy, 810
Wakabayashi, Kenta, 672
Wakeham, David, 455
Waldmann, Annika, 1005
Waljee, Akbar, 876
Walker, Alvin, 667
Walker, Annika, 916
Walker, Hannah, 153
Wallace, Michael, 915
Wallinga, Jacco, 109, 317
Waltemath, Dagmar, 114, 127, 445, 626
Walter, Sebastian, 1050, 1069, 1073, 1077
Walz, Tim, 616
Wang, Gao-yang, 684
Wang, Jifan, 735
Wang, Junfeng, 90, 285, 878, 880
Wang, Kim Zhipei, 90, 285, 880
Wang, Larry, 716
Wang, Lingjiao, 866
Wang, Ming-Dauh, 673
Wang, Shengfeng, 878
Wang, Stella, 985
Wang, Ya, 203
Wang, Yuhe, 939
Wang, Ziyuan, 810
Wang-Sattler, Rui, 891
Warnecke, Yannik, 620, 628
Warneke, Moritz, 1033
Waser, Markus, 219
Wason, James, 78, 357, 521, 1032
Wassmuth, Angelika, 301
Watanabe, Mana, 957
Watanabe, Yuya, 796
Wattar, Ahmad, 346, 908, 909
Waymyers, Sophia, 558
Weber, Frank, 383, 731
Weber, Jan, 190
Weber, Michael, 299
Weber, Philipp, 544, 1011
Weber, Sebastian, 469
Weber, Susanne, 705, 914, 1054
Weber, Thomas, 244
Weddeling, Theresa, 703
Wede, Annabell, 1001
Wedemeyer, Hanna, 942
Wee, Joseph, 210
Wegener, Silke, 931
Wehle, Maximilian, 350
Wei, Chang, 530
Weise, Annemarie, 244, 799
Weise, Julius Johannes, 449
Weiss, Helen, 961
Weiß, Manfred, 802
Weitmann, Kerstin, 761
Well, Eilin Jopp-van, 744
Wellekens, Karolien, 1000
Weller, David, 839
Wells, Molly, 659, 683, 835
Welzel, Tatjana, 609
Wen, Hui, 241
Wendlandt, Katharina, 605
Wendt, Thomas, 541, 626
Wenz, Frederik, 352
Werba, Pablo, 1030
Werle, Kim Katrin, 540
Werner, Patrick, 631
Weschenfelder, Friederike, 930

West, Joe, 974
Westermann, Lukas, 493
Westhus, Britta, 566
Westphal, Max, 131
Wettstein, Reto, 629
Weymann, Maria, 1066
White, Bethan, 661
White, Ian, 152, 153, 155, 412, 599, 702
Whittle, Rebecca, 288, 462, 642, 643, 644
Widmer, Bastian, 840
Widmer, Lukas Andreas, 469
Widmer, Michèle, 840
Widschwendter, Andreas, 593
Wiedekopf, Joshua, 536
Wiederhold, Alexander Johannes, 233
Wielga, Jenny, 703
Wiemer, Jan, 959
Wiendl, Heinz, 349, 705, 1052, 1054, 1060
Wienke, Andreas, 367, 999
Wieseler, Beate, 104
Wiesenbach, Philipp, 896
Wiesmann, Niklas, 791
Wiesner, Martin, 224, 903
Wiest, Isabella, 147, 159, 176, 178
Wijesuriya, Rushani, 433
Wild, Philipp S., 370, 875
Wilhelmi, Selina, 775
Wilken, Matthias, 104, 105
Williamson, Elizabeth, 107, 108, 361, 457, 582, 690
Wilson, Kevin, 681
Wilson, Nina, 681
Wimmesberger, Nicole, 920
Windhorst, Anita C., 1014, 1016
Winkler, Jakob, 219, 926
Winpenny, Eleanor, 201
Winter, Alfred, 622
Wit, Ernst C., 184, 188
Wittenberg, Ian, 999
Wodchis, Walter, 1026
Wodke, Judith, 114, 127
Woillard, Jean-Baptiste, 516, 717
Wojnowski, Leszek, 664
Wójtowicz, Mariusz, 1018
Wolbers, Marcel, 459
Wolf, Kathrin, 637
Wolf, Klaus-Hendrik, 604, 634
Wolf, Thomas, 449
Wolff, Robert, 637, 758
Wolfien, Markus, 621
Wolfson, Julian, 416
Wolkenstein, Andreas, 917
Wolkewitz, Martin, 107, 110, 784, 1055
Wöller, Nico, 760, 761
Wolter, Steven, 98, 99
Wong, Jamie, 855
Wood, Simon, 109
Woodward, Owen, 493
Wouters, Michel, 658
Wright, Marvin N., 513
Wu, Sihan, 196
Wudy, Stefan, 570
Wulf, Gerald, 541
Wünstel, Justyna, 398
Wüstner, Stefanie, 104, 105
Wuttke, Matthias, 571
Wynants, Laure, 90, 91, 285, 390, 646, 880
Xi, Guoren, 541
Xia, Shuang, 684, 752
Xin, Wenjing, 814
Xu, Long-Hao, 182, 311
Xu, Xingzhi, 525
Yadav, Sudha, 316
Yan, Tianjiao, 317
Yang, Chunxia, 532
Yang, Hanning, 257, 704, 1048, 1076, 1078
Yang, Jiancheng, 1069
Yasmin, Zahina, 877
Yates, Tom, 939
Yavuz, Sinan, 916
Yazici, Ayse Canan, 913