



**Statistical Thinking Across Borders:
Building Bridges and Expanding
Horizons in Clinical Biostatistics**

Conference Program & Abstract Book

**46th annual conference of the
International Society for Clinical Biostatistics**

24 - 28 August 2025, Basel Switzerland



Version: August 24, 2025

The open-source L^AT_EX template, AMCOS_booklet, used to generate this booklet is available at
https://github.com/maximelucas/AMCOS_booklet

Contents

Welcome Message	7
About the Conference	9
Committees	10
Local organizing committee	10
Scientific program committee	11
Sponsors	12
Sustainability Concept	14
ISCB award winners	15
Conference Attendance Support Program	15
Student Award Program	15
General Information	16
Conference venue	16
Arrival	17
Luggage storage	18
Catering	18
WiFi	19
Floor plan	21
BaselCard	24
List of Badge Abbreviations	25
Certificate of attendance	26
ISCB annual general meeting	27
Social Program	28
[SG] Student Gathering – Sunday evening, 24 August	29
[!] Welcome Reception – Monday evening, 25 August	31
[SA] Social Activities – Tuesday afternoon, 26 August	34
[Dinner] Conference Dinner – Wednesday evening, 27 August	53
Childcare	56
Biostats-tourism: Things to look out for	57
Conference Schedule	59
Pre-conference Courses	65
Keynote Talks	71

Contents

Invited Sessions	73
Monday, 25 August	73
When Worlds Collide: Common Methodological Themes in Meta-Analysis, Causal Inference, and Hybrid Trial Design	73
AI-Driven Multi-Omics Data Integration	76
Prediction Modelling Meets Causal Inference for Clinical Decision Making	78
Tuesday, 26 August	81
Mathematical and Statistical Modelling in the Life Sciences: Seeking Causal Explanations	81
Optimizing Efficiency in Adaptive Trial Designs	83
Wednesday, 27 August	86
Cracking Causal Questions: Estimands for Reliable and Clinically Relevant Evidence	86
Improving Replicability in Clinical Biostatistics	88
Generative AI in Clinical Research and Drug Development (BBS Invited Session)	91
Abstracts of Contributed Talks	93
Monday, 25 August	93
Efficient Analyses of Clinical Trials	93
Poster Highlights	98
Dynamic Borrowing and Basket Trials	102
Causal Inference - Dealing with Bias	108
Survival Analysis 1	114
Prediction / Prognostic Modelling 1	120
Defining Estimands for Clinical Trials	127
Adaptive and Multi-Arm Multi-Stage Trials	133
Causal Inference: Target Trial Emulation	140
Meta-Analysis 1	145
Prediction / Prognostic Modelling 2	152
Joint / Longitudinal Modelling	157
Estimands: Causal and Multiple Imputation Approaches	163
Clinical Trials and Regulatory Issues	168
Model Selection and Simulations	174
Bayesian Methods 1	179
Multi-Omics Data Integration	185
Biomarker Studies & Diagnostic Tests	191
Tuesday, 26 August	197
Clinical Trials with Longitudinal and Clustered Data	197
Bayesian Methods 2	201
Machine Learning 1	207
Infectious Disease Modelling	212
Meta Science 1	217

Contents

Design and Analysis of Trials in Rare Diseases	223
Survival and Recurrent Events in Clinical Trials	228
Alternative Estimands in Causal Inference	234
Machine Learning, Deep Learning, and AI	239
Meta-Analysis 2	245
Meta Science 2	251
Statistical Methods in Epidemiology	257
Wednesday, 27 August	263
Design and Evaluation of Clinical Trials	263
Competing Events and Multi-State Modelling	270
Machine Learning 2	277
Prediction / Prognostic Modelling 3	282
Infectious Disease and Longitudinal Modelling	288
Biomarker Studies & Mixed Topics	293
Randomization and Analysis of Clinical Trials	299
Survival Analysis 2	305
Causal Inference: Mixed Topics	311
Innovation in Oncology Dose Escalation Trials and Beyond	316
Missing Data and Imputation	322
Observational/Real-World Data 1	329
Efficient Use of Interim Analyses in Clinical Trials	335
Survival Analysis 3	341
Causal Inference in Time-Varying Settings	347
Prediction / Prognostic Modelling 4	352
Meta-Analysis 3	359
Observational/Real-World Data 2	365
Abstracts of Contributed Posters	372
Monday, 25 August	372
Monday Posters at Biozentrum	372
Monday Posters at ETH	425
Tuesday, 26 August	465
Tuesday Posters at Biozentrum	465
Tuesday Posters at ETH	515
Wednesday, 27 August	554
Wednesday Posters at Biozentrum	554
Wednesday Posters at ETH	600
Mini Symposia	643
Causal Inference for Improved Clinical Collaborations: A Practicum	643
Objective	643
Agenda	644

Contents

Early Career Biostatisticians' Day	645
Student Gathering	645
ECB Day	645
Schedule	648
Statistical Research needs to improve – on the important roles of simulation studies and guidance for analysis	649
Objective	649
Program	650
Enhancing Cancer Clinical Trials with Patient-Reported Outcomes: Insights from SISAQOL-IMI	660
Objective	660
Presenters	660
Outline	662

Welcome Message

On behalf of the Scientific Programme Committee (SPC) and the Local Organizing Committee (LOC), we are delighted to welcome you to Basel for the 46th Annual Conference of the International Society for Clinical Biostatistics (ISCB).

Nestled at the tri-border of Switzerland, Germany, and France, Basel is a city that seamlessly blends history, culture, and innovation. Its picturesque location along the Rhine River and its iconic bridges make it both a literal and symbolic venue for this year's conference. As the home of pioneering thinkers like Jakob Bernoulli, Leonhard Euler, and Paracelsus - whose contributions to probability, mathematics, and clinical medicine resonate to this day - Basel has a long tradition of advancing scientific thought. This legacy continues today as Basel is widely recognized as one of the world's leading hubs for life sciences, health research, and drug development. Many academic institutions, pharmaceutical companies and biotech firms are based here or nearby, making it the perfect venue to explore and expand the frontiers of clinical biostatistics.

The theme of this year's conference, "Statistical Thinking Across Borders: Building Bridges and Expanding Horizons in Clinical Biostatistics" reflects the importance of fostering connections - between disciplines, methodologies, and countries - to address the opportunities and challenges of today's increasingly complex healthcare and clinical research environment. With 867 in-person participants, 23 virtual attendees, and representation from 49 countries, the ISCB46 reflects the collaborative and global nature of our field.

We are pleased to offer a rich scientific program that includes 6 pre-conference courses, 2 keynote speeches from experts in Bayesian clinical trial analysis and machine learning for individualized treatments, 8 invited sessions, 4 mini-symposia and 48 contributed sessions. In total, 260 oral presentations, 30 poster flash-talks and 248 posters will showcase contributions across a wide range of biostatistical topics, including causal inference, analysis of real-world data, clinical trial design and analysis, longitudinal modeling, and time-to-event analysis. We wish to thank all the scientific contributors for shaping the ISCB46 program.

The ISCB46 also demonstrates a strong commitment to sustainability and social responsibility, emphasizing that scientific excellence can align with environmental care and the promotion of diversity and inclusion. Thanks to our sustainability sponsor BWT (Best Water Technology), all participants will receive reusable Climate Bottles to reduce plastic waste, with water dispensers provided throughout the venue. A recycling initiative for coffee cups and a zero single-use tableware policy during the conference dinner further enhance our ef-

Welcome Message

forts. In addition, Basel's efficient public transport system removes the need for conference buses and is further complemented by the free BaselCard provided to overnight guests. We are also very proud to partner with St. Josefshaus, an organization renowned for promoting diversity and inclusion by empowering individuals with disabilities. This year's conference bags were produced in their workshops, providing meaningful employment opportunities for those who might face challenges accessing the general job market. By supporting inclusive collaborations and environmentally conscious practices, ISCB46 aims to inspire a spirit of positive impact, collaboration and care.

This event would not have been possible without the dedication, expertise and tireless efforts of the 33 members of the LOC, the 20 members of the SPC and the 14 supporting reviewers. From designing a diverse and engaging program to managing the countless logistical details required to bring this conference to life, their commitment has been extraordinary. Their collaborative spirit and motivation have been the driving force behind the success of ISCB46.

In addition, we are especially grateful to our sponsors. The contributions of our institutional and industry sponsors have been crucial to the success and smooth operation of ISCB46. Last but not least, we are grateful for the trust and support of the ISCB and the Basel Biometric Society, who gave us the opportunity to host this year's meeting.

We would also like to recognize and congratulate the 16 recipients of the ISCB Conference Attendance Support Program (CASP) and the 4 recipients of the ISCB Student Conference Award (StAC) awards. The CASP award embodies ISCB's commitment to inclusivity, providing travel support to participants who might otherwise face barriers to attending, while the StAC award honors post-graduate students whose high-quality research demonstrates relevance and impact in applying statistical methods to clinical and epidemiological research. These awards reflect not only the society's dedication to fostering diversity within our community but also the pursuit of excellence in biostatistics. We are delighted to welcome these awardees to ISCB46 and celebrate their contributions to the field.

In closing, we invite you to immerse yourself in all that ISCB46 and Basel have to offer. Basel provides an inspiring backdrop for the exchanges, discussions, and collaborations that define ISCB46. May this week spark new ideas, foster connections, and broaden perspectives in our shared commitment to advancing clinical biostatistics

Vanessa Didelez (Chair of SPC)
Mouna Akacha (Co-chair of LOC)
Kaspar Rufibach (Co-chair of LOC)

About the Conference

Committees

Local organizing committee

Mouna	Akacha	Novartis	Core member	Co-chair
Kaspar	Rufibach	Merck KGaA	Core member	Co-chair
Bibiana	Blatna	Novartis	Core member	
Muriel	Buri	Roche	Core member	
Lilla	di Scala	J&J	Core member	
Eliane	Imfeld	Novartis	Core member	
Jack	Kuipers	ETH Zurich	Core member	
Roland	Marion-Gallois	BMS	Core member	
Carmen	Pasquale	Novartis	Core member	
David	Warne	Consultant Biostatistician	Core member	
Lukas	Widmer	Novartis	Core member	
Emmanuel	Zuber	Cogitamen	Core member	
Trista	Bintoro	Philip Morris	Extended member	
Jocelyn	Buisker	Novartis	Extended member	
Clélia	Cahuzac	Roche	Extended member	
Tafadzwagladys	Dhokotera	Swiss TPH & Uni Basel	Extended member	
Luisa	Eggenschwiler	Uni Basel	Extended member	
Youyou	Hu	Roche	Extended member	
Dirk	Klingbiel	BMS/Priothera	Extended member	
Baldur	Magnusson	UCB	Extended member	
Joana	Marques Barros	Idorsia	Extended member	
Antonella	Mazzei	BMS	Extended member	
Charline	Meré	BMS	Extended member	
Giusi	Moffa	Uni Basel	Extended member	
Tobias	Mütze	Novartis	Extended member	
Nikki	Rommers	University Hospital Basel	Extended member	
Amanda	Ross	Swiss TPH & Uni Basel	Extended member	
Fred	Sorenson	Cencora	Extended member	
Hong	Sun	BMS	Extended member	
Balint	Tamasi	DSM Firmenich	Extended member	
Diane	Uschner	Roche	Extended member	
Fiona	Vanobberghen	Swiss TPH & Uni Basel	Extended member	
Pierre	Verweij	Idorsia	Extended member	

Scientific program committee

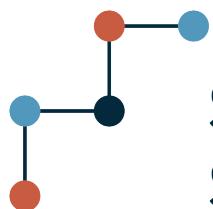
Vanessa	Didelez	Leibniz Institute for Prevention Research and Epidemiology, Uni Bremen	Chair
Mouna	Akacha	Novartis	
Niko	Beerenwinkel	ETH Zurich	
Anne-Laure	Boulesteix	Uni Munich	
Lilla	di Scala	J&J	
Krista	Fischer	Uni Tartu	
Paul	Gustafson	Uni British Columbia	
Lisa	Hampson	Novartis	
Torsten	Hothorn	Uni Zurich	
Ruth	Keogh	LSHTM	
Katherine	Lee	Murdoch Children's Research Institute	
Giusi	Moffa	Uni Basel	
Samuel	Muller	Macquarie Uni	
Daniel	Nevo	U Tel Aviv	
Aris	Perperoglou	GSK	
David	Phillippo	Uni Bristol	
Marci	Rückbeil	Sanofi	
Kaspar	Rufibach	Merck KGaA	
Kelly	Van Lancker	Uni Ghent	
Marcel	Wolbers	Roche	

Sponsors

Sustainability Sponsor



Institutional Sponsors



**Swiss National
Science Foundation**



ETH zürich

Department of Biosystems
Science and Engineering

Platinum Sponsors



Gold Sponsors



Inspired by patients.
Driven by science.



BeOne



Bristol Myers Squibb[®]



Silver Sponsors



Boehringer
Ingelheim



Springer

Bronze Sponsors



Sustainability Concept

As part of our commitment to ecological sustainability during the 46th Annual Conference of the International Society for Clinical Biostatistics, we are taking several initiatives to ensure our event is as green as possible.

We are thrilled to partner with **BWT** (Best Water Technology) - our esteemed sustainability sponsor. BWT is aligned with our sustainability vision and will generously provide each conference attendee with a **specially designed Climate Bottle**. These reusable bottles will eliminate the need for single-use plastic, contributing significantly to our plastic waste reduction goals. Attendees will have access to water dispensers throughout the venue to refill these bottles, ensuring a constant supply of fresh water while promoting sustainable practices.

Further, we are introducing a dedicated collection system for used paper coffee cups. Our catering partner has committed to recycling these cups, ensuring they are processed responsibly. We kindly ask all attendees to support this initiative by disposing of their used coffee cups in the clearly marked recycling bins located throughout the venue.

We are proud to share that our conference dinner venue partner, Restaurant Seegarten, fully supports our sustainability goals. They have confirmed that no disposable tableware will be used during the event. All dining materials will be either reusable or sustainably sourced, reinforcing our shared commitment to reducing waste and minimizing environmental impact throughout ISCB 2025.

Moreover, in an effort to reduce emission and carbon footprint, we have chosen not to organize buses for social activities. Instead, we encourage all attendees to utilize Basel's excellent public transportation system. When you stay overnight in Basel, you will receive the BaselCard and benefit from free travel on public transport. The booking confirmation of your hotel or Airbnb in Basel serves as a free ticket for the transfer from the airport or train station to your hotel.

Lastly, all lunch boxes made available during the conference will be fully vegan, while we ensure to accommodate other dietary requirements as needed.

We are dedicated to making ISCB46 an event to be remembered, not just for its scientific contributions, but for its dedication to ecological sustainability as well.

ISCB award winners

Conference Attendance Support Program

Hongqiu	Gu	Beijing Tiantan Hospital, China
Hollie	Hughes	University of Liverpool, United Kingdom
Nia	Kang	McGill University, Canada
Nour	Kanso	King's College London, United Kingdom
Anita	Kerubo	Aga Khan University, Kenya
Kuldeep	Kumar Sharma	National Institute of Mental Health and Neurosciences, India
Shambhavi	Mishra	University of Lucknow, India
Graceful	Mulenga	Botswana Harvard Health Partnership, Botswana
Brendah	Nansereko	London School of Hygiene and Tropical Medicine, UK
Qingyang	Shi	University of Groningen, Netherlands
Maryam	Shojaei Shahrokhabadi	Hasselt University, Belgium
Vibha	Srichand	Manipal Academy of Higher Education, India
Shashank	Tripathi	University College of Medical Sciences New Delhi, India
Yiling	Zhou	University of Groningen, Netherlands

Student Award Program

Muhitul	Alam	University of Dhaka, Bangladesh
Kanella	Panagiotopoulou	Université Paris Cité, INSERM, France

General Information

Conference venue

The ISCB46 is held in the **Biozentrum of the University of Basel** and the **Department of Biosystems Science and Engineering (D-BSSE) of the ETH Zurich**, both located in the heart of Basel, in close vicinity to the Rhine. Both facilities host some of the leading life sciences institutes in the world.

Biozentrum, University of Basel

Spitalstrasse 41
CH - 4056 Basel
Switzerland

ETH Zurich | D-BSSE

Klingelbergstrasse 48
CH - 4056 Basel
Switzerland

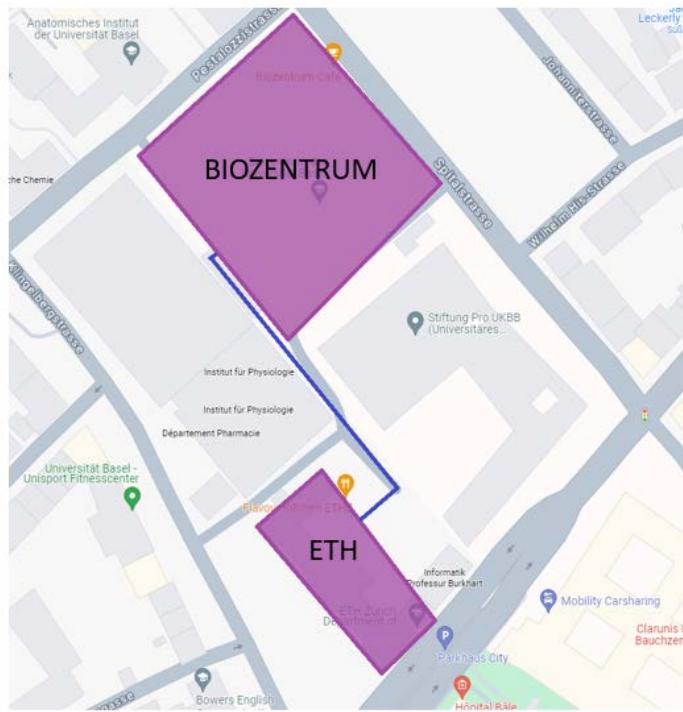


(a) Biozentrum



(b) ETH - D-BSSE

General Information



Location of the two conference venues, 2 minutes walk apart

Arrival

Upon arrival at the conference venues, participants are kindly asked to proceed to the registration desk located in the entrance hall at the U1 (-1) level of the Biozentrum, University of Basel.

Registration Desk Opening Hours:
Sunday 24 to Thursday 28 August 2025 08:00-17:00

At the registration desk, you will receive your conference badge and welcome materials. Please make sure to wear your badge visibly at all times throughout the conference, as it grants access to all sessions and venues, as well as to the catering stations.

For assistance with directions or logistical questions, staff members will be available at the information desk next to the registration desks throughout the entire conference.

Luggage storage

For your convenience, a designated luggage storage room is available on the U1 level of the Biozentrum. Please note that this room is not locked, and storage is provided at your own risk. We kindly recommend that you do not leave valuables or important items unattended.

Catering

To support a sustainable and inclusive conference experience, we have arranged the following catering services for all participants:

Water Bottles and Water Stations

Each participant will receive a reusable water bottle upon registration on Sunday, or at the bottle handout in the ETH BSSE building on Monday. Please bring your bottle with you throughout the conference as no replacements will be provided.

Water stations offering both still and sparkling water are located at four points across the Biozentrum and ETH BSSE buildings. Furthermore, tap water is drinkable and delicious in Switzerland. We encourage all attendees to make use of these refill stations regularly. Kindly note: Only closed water bottles are permitted inside the lecture halls.

Coffee and Tea

Complimentary coffee and tea will be available:

- Before the first session each day
- During designated coffee breaks
- During the lunch break

The catering stations are marked on the floor plans and will also be signposted on site. During the morning and afternoon coffee breaks, fruits and snacks are planned. All participants with dietary restrictions (e.g., lactose intolerance, egg allergy, gluten free diet) can go to the coffee station on the atrium of the ETH BSSE building for their snack.

Lunches

On Sunday, a warm lunch will be served at the Biozentrum for all course participants.

From Monday through Thursday, lunch bags will be distributed at clearly marked distribution points in both venues (see floor plans and signage on site).

The default lunch option is vegan, ensuring accessibility for most participants, including those with allergies (e.g., lactose intolerance, allergies to eggs, seafood, or shellfish). If you indicated other specific dietary requirements (e.g., gluten-free, nut allergies), your lunch bag will be available for pickup at the ETH BSSE building (atrium catering station). In case of special dietary restrictions, you have a sticker on your badge that matches the sticker on your special lunch bag (e.g., avocado, chili, pineapple, garlic), but the menu was chosen to accommodate most participants with dietary restrictions. All allergens will be highlighted so you can always double check. Note that no distinction is made between participants allergic to a specific type of nut or participants allergic to all nuts.

WiFi

Free WiFi is available for all participants at both conference venues. All attendees with an academic account, can connect to the eduroam network that is available in both venues. Instructions to the public WiFi:

Biozentrum - University of Basel

Connect to the WiFi network "unibas-visitor" and open your browser. It will redirect you to the welcome page. If this fails, try opening a non-secure site (e.g. <http://unibas.ch>). Click "Continue here" to reach the Monzoon login page and select your preferred language (English available). If you already have a code, log in with your mobile number and access code. If not, click on "If you don't have a code yet, please click here to register" and enter your mobile number. You'll receive an access code via SMS free of charge. Enter the code and click "Connect".

Alternative access: If you don't receive an SMS or don't have a mobile phone, call +41 (0)43 500 3456 from a Swiss landline or international mobile. A spoken code will be read to you (local call charges apply). For support, contact Monzoon's infoline: 0800 666 966 (available 24/7).

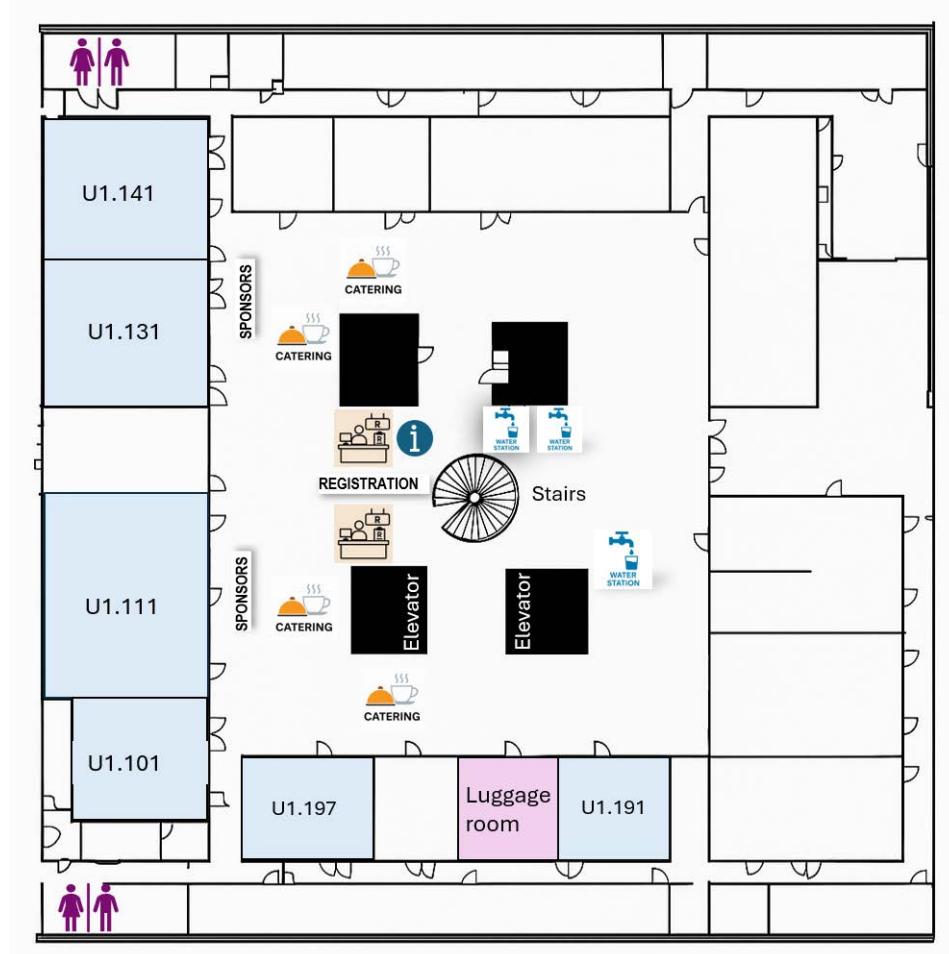
BSSE Building - ETH Zurich

Connect to the WiFi network "ETH-Visitor" and open your browser. Also here, you should register with your mobile phone number after which you receive a login code via SMS. Enter the code to access the internet. Detailed instructions can be found here:

<https://unlimited.ethz.ch/spaces/itkb/pages/276478014/ETH-Visitors>

Floor plan

Biozentrum - Main conference floor U1 (level -1)

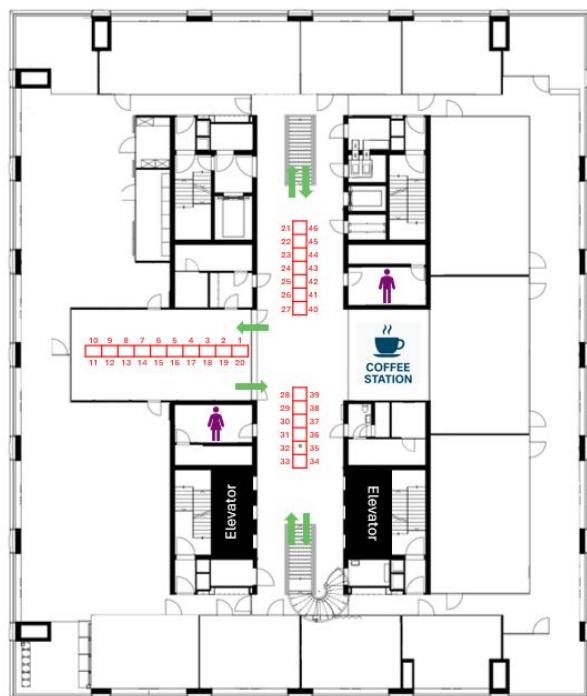


ETH - Main conference floor E (ground floor level)

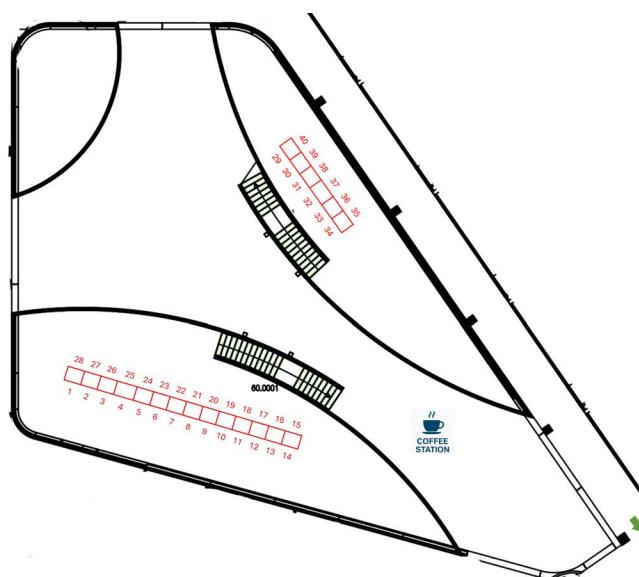


Poster exhibition:

- Biozentrum level 2
- ETH level D (level -1, below atrium)



(a) Poster area Biozentrum (BioZ posters)



(b) Poster area ETH (ETH posters)

BaselCard

The **BaselCard** is offered as a free bonus with every booking at a Basel hotel, hostel, bed and breakfast or apartment. You will receive your personal guest card when you check in, and you can also load it to your smartphone as a web app.

Some of the main advantages of the BaselCard at a glance:

- Free travel on all public transport in Basel throughout your entire stay, including your initial journey to your hotel from EuroAirport Basel-Mulhouse-Freiburg or one of the three train stations in Basel (Basel SBB train station, Basel Bad Bhf and Basel SNCF). For your initial journey the booking confirmation of your accommodation serves as a free ticket for this transfer.
- 50% discount
 - on museum entrances and the entrance to Basel Zoo
 - on the public walking tour of the Basel Old Town and the tour on the sightseeing bus
 - and many additional touristic offers

For more information, see www.basel.com/en/baselcard.

List of Badge Abbreviations

- **Pre-Conference courses**, Sunday afternoon, 24 AUG 2025

C1 Full Day Course 1: Targeted Learning

C2 Full Day Course 2: The Design of Simulation Studies

C3 Full Day Course 3: Network Meta-Analysis: From Key Concepts to Advanced Methods

C4 Full Day Course 4: Good Software Engineering Practice for R Packages

CM Half Day Course (morning): Bayesian Methods for Precision Medicine

CA Half Day Course (afternoon): Multi-State Models: Theory, Applications and New Developments

 **Welcome Reception**, Monday evening, 25 AUG 2025

- **Social Activities**, Tuesday afternoon, 26 AUG 2025

SA1 Guided Tour: Street Art Guided tour of Basel

SA2 Guided Tour: Stories of Basel's Old Town

SA3 Guided Tour: Novartis Campus Architecture

SA4 Confiserie Beschle – Chocolate factory visit and create your own chocolate bar

SA5 Adventure-filled Scavenger Hunt in Basel

SA6 Crossbow event – William Tell, the central figure of the Swiss Confederacy

 **Conference Dinner**, Wednesday evening, 27 AUG 2025

- **Mini-Symposia**, Thursday, 28 AUG 2025

MS1 Mini-Symposium 1: Causal Inference for Improved Clinical Collaborations: A Practicum

General Information

MS2 Mini-Symposium 2: Early Career Biostatisticians' (ECB) Day

MS3 Mini-Symposium 3: Statistical Research needs to improve – on the important roles of simulation studies and guidance for analysis

MS4 Mini-Symposium 4: Enhancing Cancer Clinical Trials with Patient-Reported Outcomes: Insights from SISAQOL-IMI

Certificate of attendance

By default, no certificate of attendance will be provided. If you need one please refer to the information desk. You can leave your name and email there and will get a certificate of attendance electronically.

ISCB annual general meeting

The International Society for Clinical Biostatistics (ISCB) is holding its 2025 Annual General Meeting on **Wednesday, 27 August 2025 at 12.15 - 13.30 CEST in Biozentrum U1.111** at the ISCB46 in Basel, Switzerland.

The ISCB will be opening attendance and voting during the AGM to its entire pool of active membership. Registered delegates of the ISCB46 Conference are invited to participate physically. ISCB members not registered to the Conference can still partake in the works of the AGM via online voting.

Please find the 2025 AGM agenda and access details as well as instructions for participation and voting in the e-mail sent by the ISCB office. Reports and Minutes for the current and past AGM can be accessed through the Members Area at <https://www.iscb.international/agm-updates/>

We invite you to participate actively in this important procedure for our Society.

Social Program

Information on the social program organized for the ISCB 2025 conference is summarized below.

The social program is composed of the Student Gathering (taking place on Sunday evening, 24 August), the Welcome Reception (taking place on Monday evening, 25 August), the Social Activities (taking place on Tuesday afternoon, 26 August) and the Conference Dinner (taking place on Wednesday evening, 27 August).

Please be aware that only those attendees who have registered for specific events at the time of conference registration can attend those events. Check your badge for abbreviations to see which events you have signed up for (refer to the "List of Badge Abbreviations" section in this booklet). For example, if your badge displays "SG SA2 " it means you are registered for the Student Gathering (SG), Guided Tour on "Stories of Basel's Old Town" (SA2), and the Conference Dinner (). You will not be able to attend the Welcome Reception or any other social activities offered on Tuesday afternoon. If you remain uncertain, please contact any member of the organizing committee.

Please prioritize walking or using public transportation to reach the various locations. Instructions for each transportation mode are provided below. Note that parking spaces are unavailable at most locations. Transportation to the various locations is not covered by the conference registration fees and must be paid for separately. If you booked an accommodation in Basel, **the BaselCard** offers you free travel on all public transport in Basel throughout your entire stay.

[SG] Student Gathering – Sunday evening, 24 August

Our annual networking event will be held in the Markthalle under the large domed roof built in 1929, offering a welcoming and unique atmosphere. It's the ideal way for students and early career biostatisticians to kick off the conference! Attendees will have plenty of opportunities to network and connect with fellow statisticians in a relaxed setting over an *apéro*.

Start time / Duration: 18:00 - 20:00 / 6 - 8 pm CET.

Location: Markthalle, Steinentorberg 20, 4051 Basel.



How to get there:

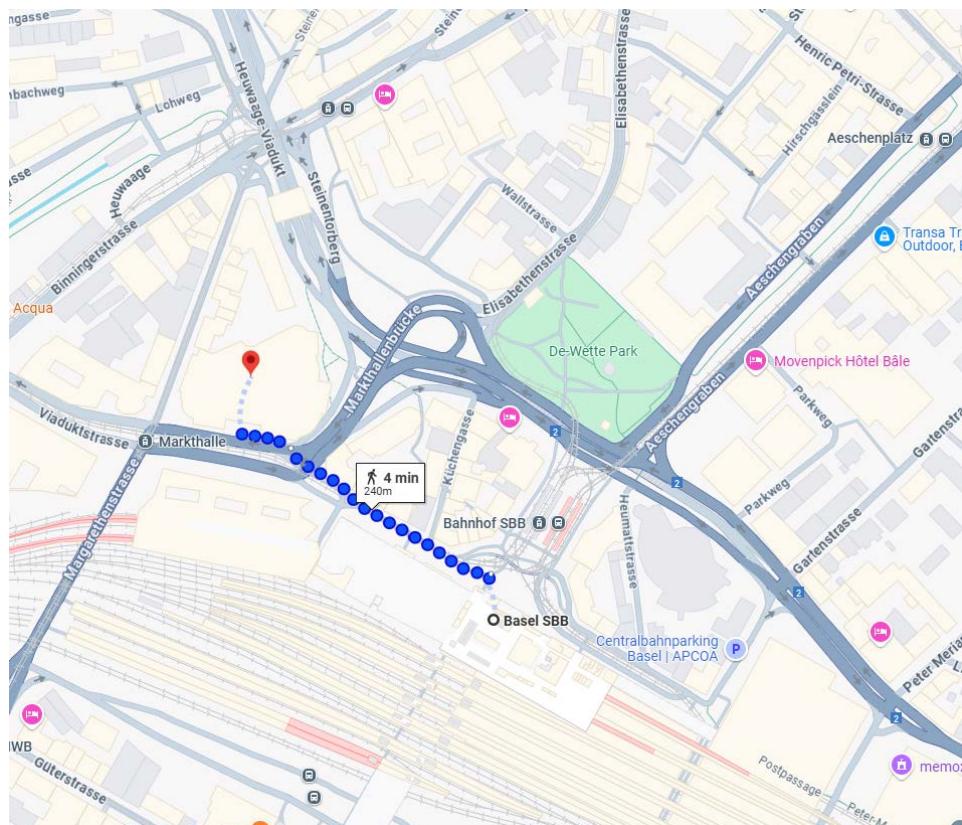
The simplest option is to head to the main train station (*Bahnhof SBB*), which is easily accessible by numerous trams and buses from various parts of the city. From there, you can walk to Markthalle, located in close proximity:

- From *Bahnhof SBB*, head west on Centralbahnstrasse (go on your right when facing the train station, see the map below) for approximately 4 minutes (~250 meters).
- You will see the Markthalle building on the right-hand side of the road (written on the building).

If you wish to join Markthalle from Biozentrum Basel (Spitalstrasse 41), see the instructions [here](#). You will use Bus 30 from *Kinderspital UKBB* [[timetable](#)].

Additional information:

- The event will start promptly as scheduled at 18:00.
- Attendance at the Student Gathering is limited to registered participants only.
- Please ensure you arrive at the meeting point at least 15-20 minutes before the start time to allow enough time to register. If you have not picked up your badge yet, you will be able to do so. Please note that there will not be any possibility to register for late arrival after 18:30.
- Each participant will receive a voucher for two drinks of their choice. Please be aware that any additional drinks will be at your own expense.
- The event is planned to finish around 20:00 and snacks will be provided. If you wish to continue the evening afterward, you can choose from the **many fine food stands available at Markthalle** for dinner. Please note that dinner is not part of the conference and will be at your own expense.



[] Welcome Reception – Monday evening, 25 August

The Welcome Reception is a social event marking the conclusion of the first day of the conference. It will take place in the courtyard of the Public Record Office. This event provides an excellent opportunity to connect and network with fellow statisticians attending the conference.

For security reasons, attendance at the Welcome Reception is limited to 250 participants, with entry allocated on a first-come, first-served basis at the time of conference registration. **Please bring along your badge** displaying the abbreviation  to ensure your access to the Welcome Reception venue.

Time: 18:00 to 20:30

Location: In the courtyard of the Public Record Office behind the Basel parliament building (Rathaus), Martinsgasse 2, 4051 Basel. **Please note that the entrance to the courtyard of the Public Record Office is located in Martinsgasse behind the parliament, and cannot be accessed by the main entrance of the parliament, located on Marktplatz.**

You can either walk or take a bus to join the meeting point (see detailed instructions and map below).



Basel parliament building (Rathaus). This is NOT the main entrance to the Welcome Reception. See detailed maps below.

How to get there on foot:

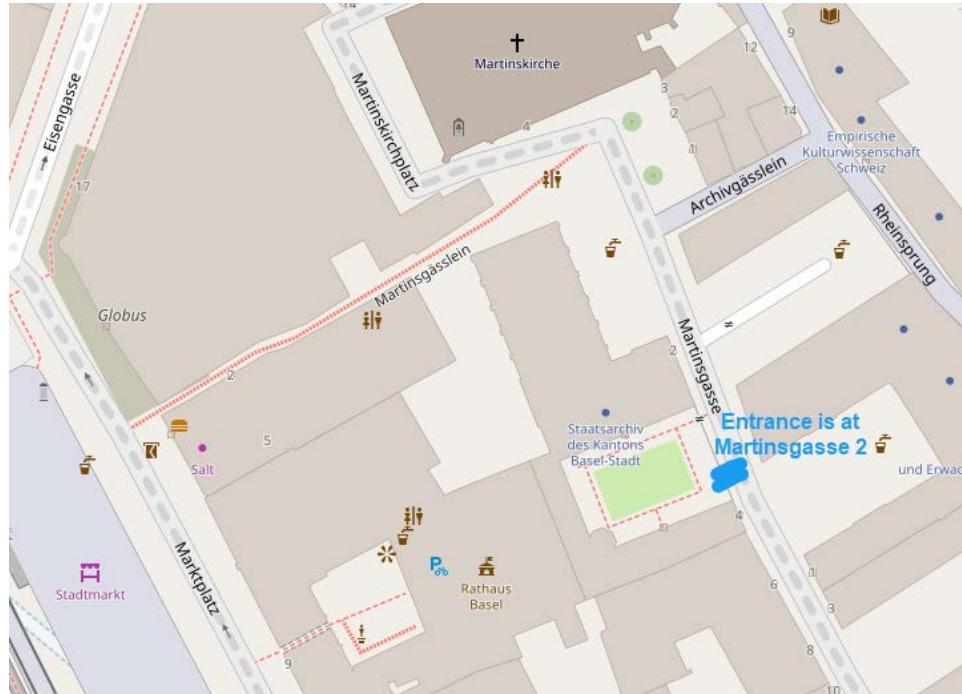
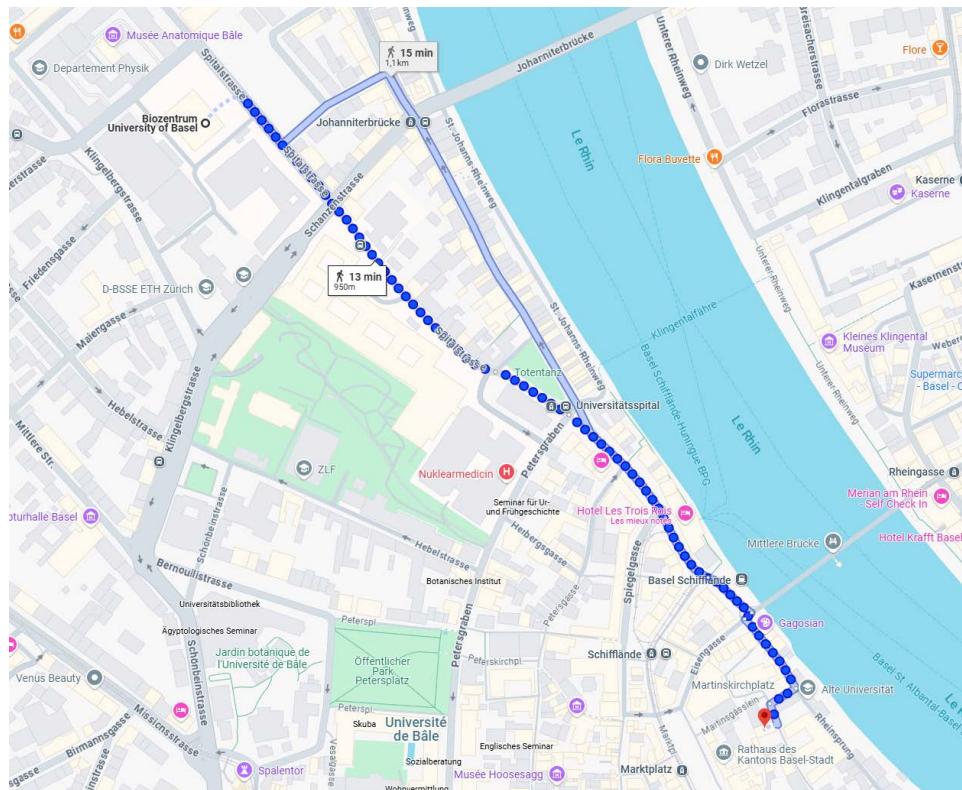
- The simplest option is to walk from the Biozentrum (approximately 15 minutes and 1 km, as indicated on the map below).

How to get there by public transport:

- From Biozentrum University of Basel, walk east on Spitalstrasse to reach *Kinderspital UKBB* bus stop at the crossing between Spitalstrasse and Schanzenstrasse (approximately 3 minutes, 150 meters).
- Take Bus 31 direction *Riehen, Otto Wenk-Platz* or Bus 33 direction *Basel, Schiffbrücke* or Bus 36 direction *Basel, Schiffbrücke* or Bus 38 direction *Wylen, Siedlung* and get off at *Basel, Schiffbrücke* (approximately 2 minutes)
- From there, walk up towards Rheinsprung (on your right when facing the Rhine river),

General Information

turn right on Archivgässlein and then left on Martinsgasse.



[SA] Social Activities – Tuesday afternoon, 26 August

Social activities are scheduled for Tuesday afternoon. Please note that the starting times for the activities may vary. Further details are provided below for each activity.

[SA1] Guided Tour: Street Art Guided tour of Basel

The tri-border area around Basel is regarded as an unrivalled street-art Mecca. Around 100 artists are active in Basel's public sphere and transform the city into a giant open-air museum with works on every corner.

On this guided group tour, you will discover colourful, urban Basel and hear the stories behind the artworks. Your guide will show you the impressive works of local artists and take you to see outstanding pieces by internationally renowned.



Pictures are taken from [Artstuebli, Basel](#).

Duration: Approximately 1.5 hours.

Language: English.

Start time: 14:00 / 2pm CET (**Be sure to arrive at the meeting point at least 15 minutes prior to the start time. The activity will start promptly as scheduled.**)

Meeting Point: In front of the “Artstübli” (located in the same building as the Markthalle food corner), Steinentorberg 28, 4051 Basel.

You can either walk or take a bus to join the meeting point (see detailed instructions and map below).



Entrance of the "Artstübl".

Additional information:

- Attendance is limited to registered participants only.
- The guided tour is conducted exclusively outdoors, so we recommend wearing sturdy and comfortable shoes.
- The tour is planned to end at Spalenberg, in the old town of Basel.

How to get there by public transport:

- From Biozentrum University of Basel, walk east on Spitalstrasse to reach *Kinderspital UKBB* bus stop at the crossing between Spitalstrasse and Schanzenstrasse (approximately 3 minutes, 150 meters).
- Take Bus 30 in the direction of *Basel, Bahnhof SBB* (main train station) and get off the bus at *Bahnhof SBB* station (approximately 8 minutes). The bus will stop in front of the train station.
- From here, head west on Centralbahnstrasse (go on your right when facing the train station, on the street indicated with a yellow arrow on the map below) for approximately 4 minutes (270 meters). You will see the Markthalle building on the right-hand side of the road (written on the building), where the "Artstübl" is located. The entrance for the "Artstübl" is located on the right-hand side of the building (when facing it from the train station) when going down on Steinentorberg.

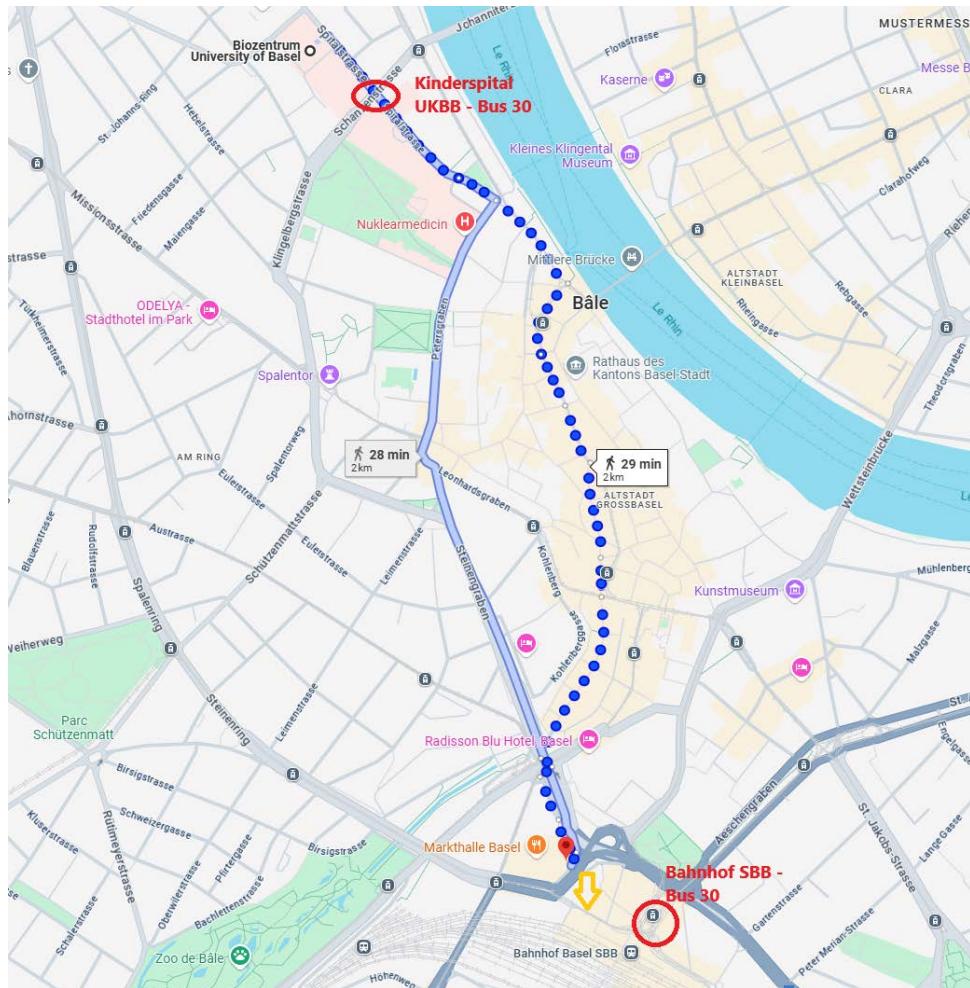
How to get there on foot:

- A 29-minute walk (approximately 2 km) from the Biozentrum University of Basel, as

General Information

indicated on the map below. We recommend this route through the city, primarily along pedestrian streets, rather than the alternative 28-minute route, also shown on the map, which follows a busier street with some traffic.

- When you arrive to the Markthalle building (the name is written on it, opposite entrance as if you take the bus), the entrance for the "Artstüblí" is located on the left-side of the building when going up on Steinentorberg.



[SA2] Guided Tour: Stories of Basel's Old Town

This traditional tour offers visitors and locals alike a close-up view of Basel. We will guide you through narrow alleyways and secluded squares, past impressive buildings and lively and vivid cityscapes, experiencing what has shaped the city and given it its unique character.



Pictures are taken from [This is Basel!](#).

Duration: Approximately 1.5 hours.

Language: English.

Start time: 14:00 / 2pm CET (**Be sure to arrive at the meeting point at least 15 minutes prior to the start time. The activity will start promptly as scheduled.**)

Meeting Point: Tinguely Fountain ("Tinguely Brunnen") by the Theatre Basel, 4051 Basel. You can either walk or take a tram to join the meeting point (see detailed instructions and map below).



Tinguely Fountain. Picture is taken from [This is Basel!](#).

Additional information:

- Attendance is limited to registered participants only.
- The guided tour is conducted exclusively outdoors, so we recommend wearing sturdy and comfortable shoes.
- The tour is planned to end at Spalenberg, in the old town of Basel.

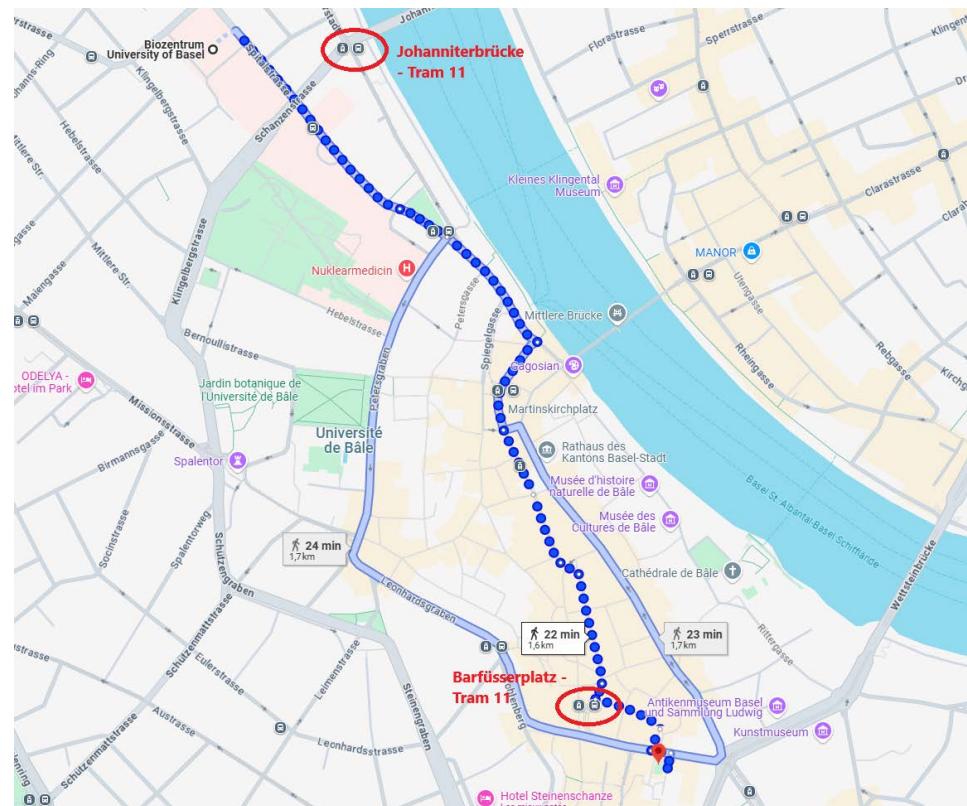
How to get there by public transport:

- From Biozentrum University of Basel, Walk southeast on Spitalstrasse toward Wilhelm His-Strasse to reach *Johanniterbrücke* tram stop (approximately 4 minutes, 260 meters).
- Take Tram 11 in the direction of *Aesch BL, Dorf* and get off the tram at *Barfüsserplatz* (approximately 7 minutes).
- Head south on *Barfüsserplatz* toward Steinenberg for approximately 2 minutes (210 meters) to reach the Tinguely fountain in front of the Theater.

How to get there on foot:

- From Biozentrum University of Basel, approximately 22-23 minutes walk through the city (mostly car-free streets) as indicated on the map below.

General Information



[SA3] Guided Tour: Novartis Campus Architecture



Pictures are taken from the [Novartis Media Library](#).

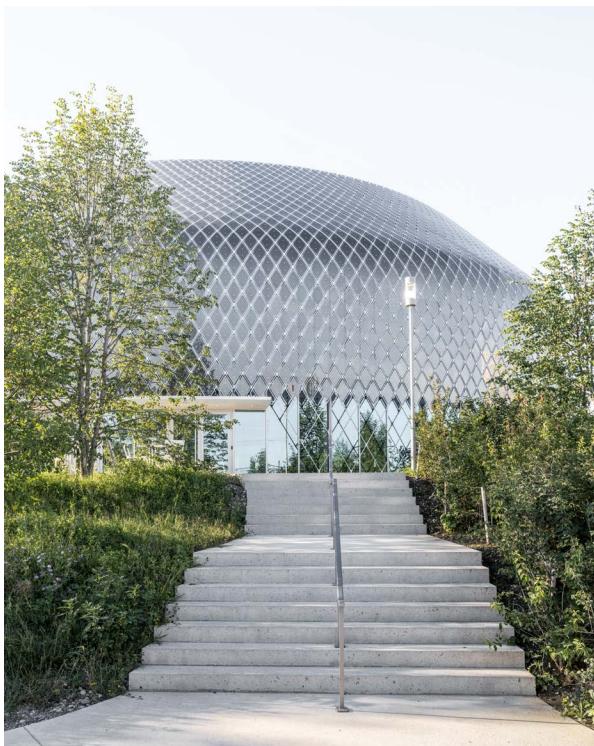
The Novartis Campus in Basel is the global headquarters of Novartis. On this tour, you will get to see the outsides of impressive office and research buildings designed by famous architects (Diener & Diener, Frank O. Gehry and Herzog & de Meuron, to name some of them). The walk will also take you through grand parks and past impressive sculptures, and you will discover the architect Vittorio Magnago Lampugnani's plan to enable dialogue and collaboration through architecture and spatial design. Subject to availability, the tour includes a visit to the "Wonders of Medicine" exhibition in the Novartis Pavillon, with its four interactive exhibition areas that explore life, disease, the history of medicine and the future of healthcare.

Duration: Approximately 1.5 hours.

Language: English.

Start time: 14:00 / 2pm CET (**Be sure to arrive at the meeting point at least 15 minutes prior to the start time. The activity will start promptly as scheduled.**)

Meeting Point: In front of the Novartis Pavillon, St. Johanns-Hafen-Weg, 5, 4056 Basel. You can either walk or take a tram to join the meeting point (see detailed instructions and map below).



Main entrance of the Novartis Pavillon. Picture is taken from the [Novartis Media Library](#).

Additional information:

- Attendance is limited to registered participants only.
- The guided tour is conducted exclusively outdoors, so we recommend wearing sturdy and comfortable shoes.
- Smoking is not permitted on the Novartis Campus.

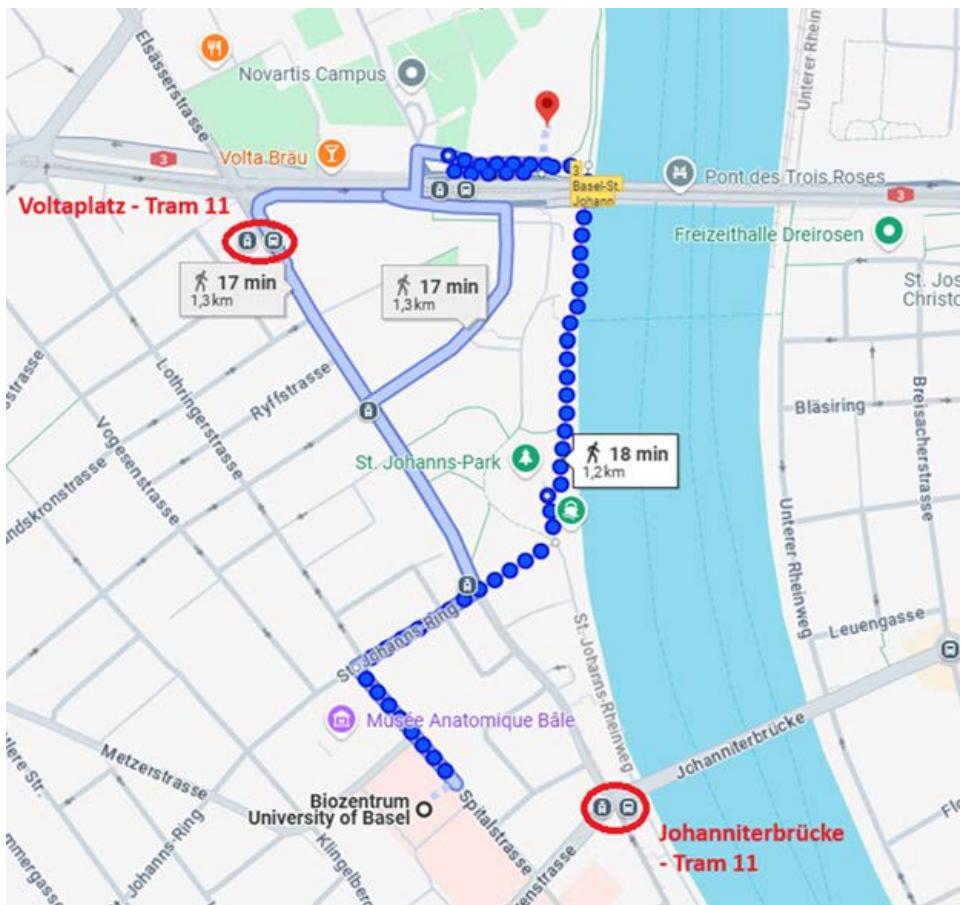
How to get there by public transport:

- From Biozentrum University of Basel, Walk southeast on Spitalstrasse toward Wilhelm His-Strasse, to reach *Johanniterbrücke* tram stop (4 minutes).
- Take Tram 11 in the direction of *St-Louis Grenze* and get off the tram at *Voltasplatz* (approximately 3 minutes).
- Head north and then right on Voltastrasse for approximately 5 minutes (450 meters), to reach the Novartis Pavillon located at the entrance of the Novartis Campus. The Novartis Pavillon is the round-shaped building covered with diamond-shaped organic photovoltaic panels located on the right when facing the Novartis Campus gates.

How to get there on foot:

General Information

- From Biozentrum University of Basel, approximately 18 minutes walk (1.2km, mostly car-free streets) as indicated on the map below. We recommend choosing this walking option that follows a promenade along the Rhine river, over the alternative routes displayed on the map.



[SA4] Confiserie Beschle – Chocolate factory visit and create your own chocolate bar

Embark on a journey through the world of chocolate and learn how professionals sample and craft this most indulgent treat. Discover the process of growing and processing cocoa beans, and gain insight into how chocolate is made and refined. Plus, enjoy the opportunity to taste different types of chocolate and, to top it all off, create your own chocolate bar.



Pictures are provided by [Beschle, Basel](#).

Duration: Approximately 2 hours.

Language: English.

Start time: 14:00 / 2pm CET (**Be sure to arrive at the meeting point at least 15 minutes prior to the start time. The activity will start promptly as scheduled.**)

Meeting Point: Confiserie Beschle, Clarastrasse 4, Basel. **Please note that Beschle has several shops across the city. The activity will take place in the one located in Clara neighborhood, on the right bank of the Rhine.**

You can either walk or take a bus to join the meeting point (see detailed instructions and map below).



Entrance of confiserie Beschle on Clarastrasse.

How to get there by public transport:

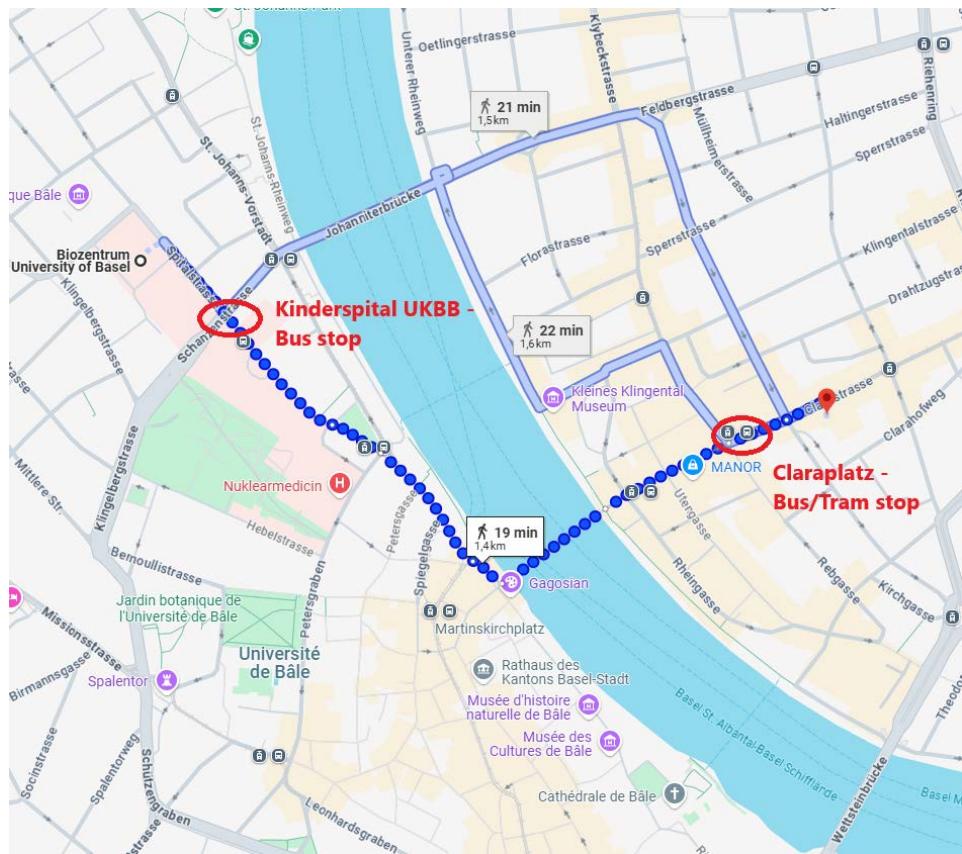
- From Biozentrum University of Basel, walk east on Spitalstrasse to reach *Kinderspital UKBB* bus stop at the crossing between Spitalstrasse and Schanzenstrasse (approximately 3 minutes, 150 meters).
- Take Bus 38 in the direction of *Wylhen, Siedlung* or Bus 33 in the direction of *Claraplatz* and get off the bus at *Claraplatz* station (approximately 5 minutes).
- Head north on Clarastrasse for approximately 1 minute (140 meters) to reach the Confiserie Beschle located on the right-hand side of the road.

Please be aware that there are various connections available to reach the event location from the Biozentrum. The most straightforward route with minimal walking is to take Bus 38/33 as described below. Alternatively, you can take Bus 36 from *Kinderspital UKBB* to Tram 6 (change at *Shifflände*), or walk towards the city center (*Shifflände, Mittlere Brücke*) and choose from the numerous trams and buses available.

How to get there on foot:

- From Biozentrum University of Basel, approximately 19 minutes walk (1.4km, mostly car-free streets) as indicated on the map below.

General Information



[SA5] Adventure-filled Scavenger Hunt in Basel

Head out to the prettiest places in the city and solve thrilling riddles. Discover hidden architectural gems and see long-known things with new eyes. An iPad guides a group of 6 attendees through the adventure, and the riddle bag holds further utensils. Additionally to difficulty and duration, there are several further customisation options to create the perfect experience for you.



Pictures are taken from [This is Basel!](#).

Duration: Approximately 2 hours.

Language: English.

Start time: 14:00 / 2pm CET (**Be sure to arrive at the meeting point at least 15 minutes prior to the start time. The activity will start promptly as scheduled.**)

Meeting Point: Breakout Basel, Steinenvorstadt 13, 4051 Basel.

You can either walk or take a tram to join the meeting point (see detailed instructions and map below).

General Information



Meeting point at Breakout Basel on Steinenvorstadt.

Additional information:

- Attendance is limited to registered participants only.
- The activity is conducted exclusively outdoors, so we recommend wearing sturdy and comfortable shoes.

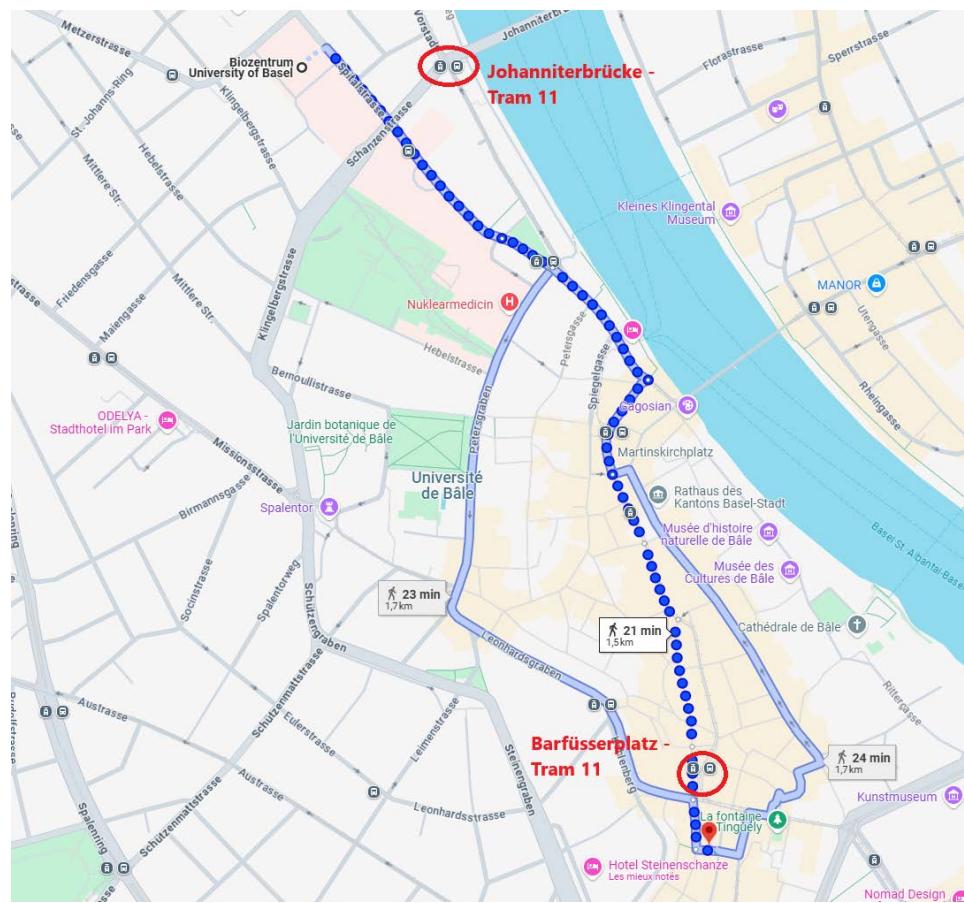
How to get there by public transport:

- From Biozentrum University of Basel, Walk southeast on Spitalstrasse toward Wilhelm His-Strasse to reach *Johanniterbrücke* tram stop (approximately 4 minutes, 260 meters).
- Take Tram 11 in the direction of *Aesch BL, Dorf* and get off the tram at *Barfüsserplatz* (approximately 7 minutes).
- Walk on Steinenvorstadt for approximately 2 minutes (200 meters) to reach the event location (located on the left-hand side of the street, in the same building as 'Mr. Pickwick Pub').

How to get there on foot:

- From Biozentrum University of Basel, approximately 21 minutes walk (1.5 km, mostly car-free streets) as indicated on the map below.

General Information



[SA6] Crossbow event – William Tell, the central figure of the Swiss Confederacy

Have you ever wondered what it feels like to be [William Tell](#)? Learn from professionals on how to use a crossbow and rediscover this traditional Swiss sport. Will your arrow find the apple's core?



Pictures are taken from [asevent GmbH](#).

Duration: Approximately 2 to 2.5 hours.

Language: English.

Start time: 14:30 / 2.30pm CET (**Be sure to arrive at the meeting point at least 15 minutes prior to the start time. The activity will start promptly as scheduled.**)

Meeting Point: Schützenhaus, Schützenstrasse 19, Reinach (BL).

The activity will take place in Reinach, near Basel, so there is no walking option to the event location. You can take a tram to easily get to the meeting point (see detailed instructions and maps below).

General Information



Meeting point: Schützenhaus in Reinach (BL). Picture is taken from the homepage of the Armbrustschützenverein Reinach - Birseck.

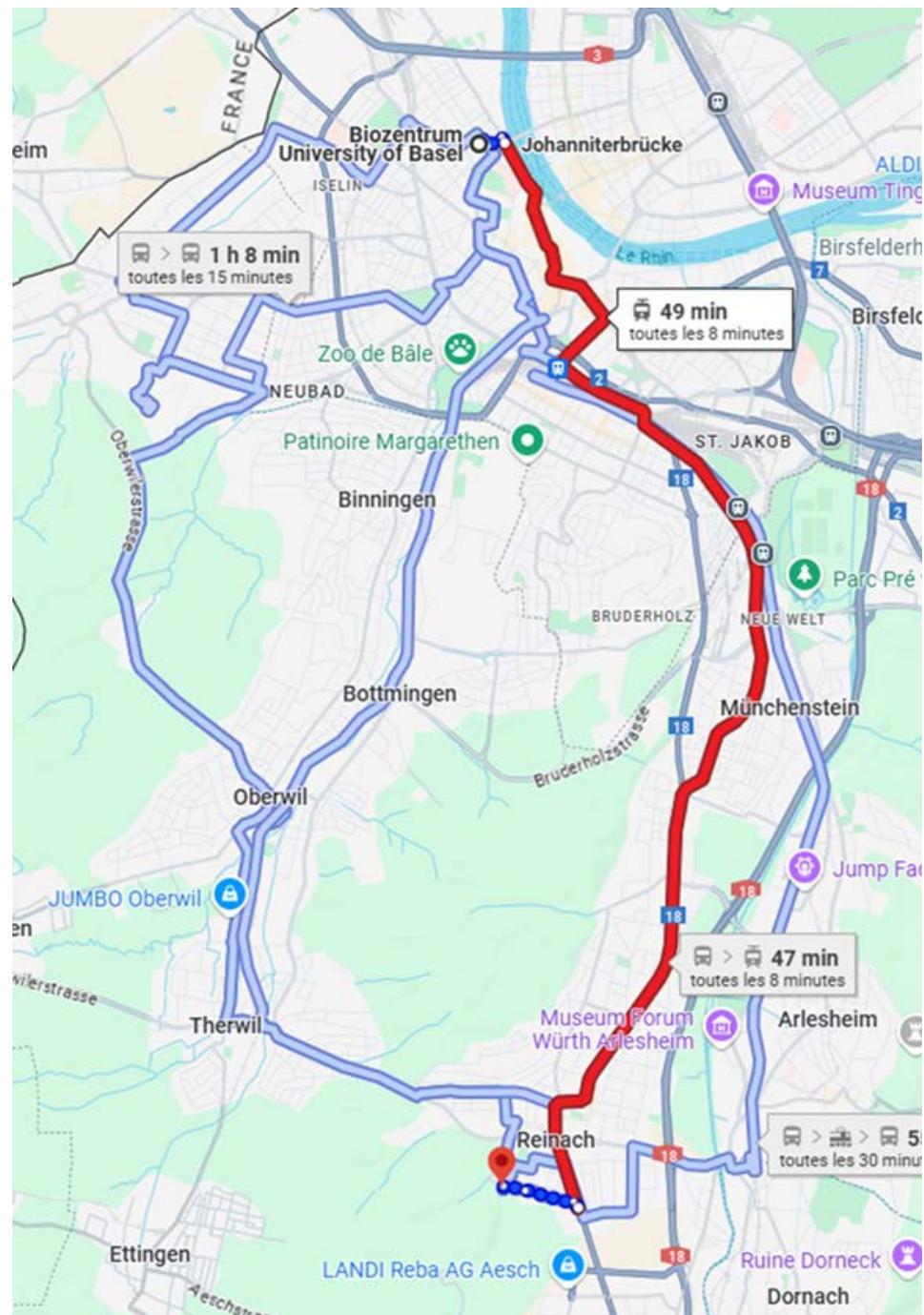
Additional information:

- Attendance is limited to registered participants only.
- We recommend wearing comfortable clothes and shoes.
- There are options to buy beverage at the location (coffee ☕, beer 🍺). Please bring some cash with you as credit cards or Twint are not accepted.

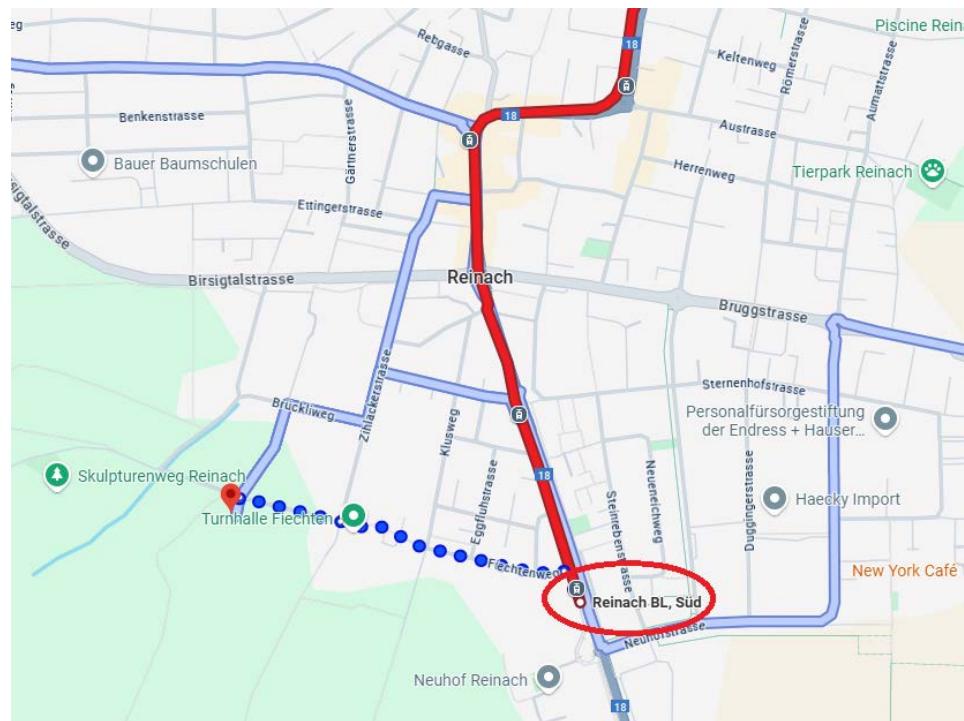
How to get there by public transport:

- From Biozentrum University of Basel, Walk southeast on Spitalstrasse toward Wilhelm His-Strasse to reach *Johanniterbrücke* tram stop (approximately 4 minutes, 260 meters).
- Take Tram 11 in the direction of *Aesch BL, Dorf* and get off the tram at *Reinach BL, Süd* (approximately 36 minutes).
- Head north and take the turn left on Fichtenweg. Walk for approximately 9 minutes (650 meters) to reach the Schützenhaus in Reinach (BL), as indicated on the map below.

General Information



General Information



[] Conference Dinner – Wednesday evening, 27 August

Enjoy a diverse selection of local and international food options at our conference dinner, featuring food booths, a relaxed atmosphere, and complimentary drinks (including beer and wine). Later in the evening, the local band "Trigger Sixx" playing some live music for us.

This is the perfect opportunity to connect with fellow attendees and speakers in a more informal and relaxed setting. We look forward to sharing this memorable evening with you!

Start time: 18:00 / 6pm CET

Meeting Point: Restaurant Seegarten, Rainstrasse 6, 4142 Münchenstein

The venue for the Conference Dinner is located slightly outside the city, so there is no walking option to the location. You can take a tram to easily get to the meeting point (see detailed instructions and maps below).

Additional Information:

- Attendance at the Conference Dinner is limited to registered participants.
- Entrance starts at 18:00 / 6pm CET and drinks will be served.
- The venue includes both indoor and outdoor spaces (with a covered area), so please bring warm clothing for the late evening.
- There is no formal dress code for the conference dinner. Feel free to dress comfortably!
- **Please bring along your badge** displaying the abbreviation  to ensure your access to the Conference Dinner venue.
- If you have any special dietary requirements or are unsure, please feel free to ask the staff at Restaurant Seegarten.

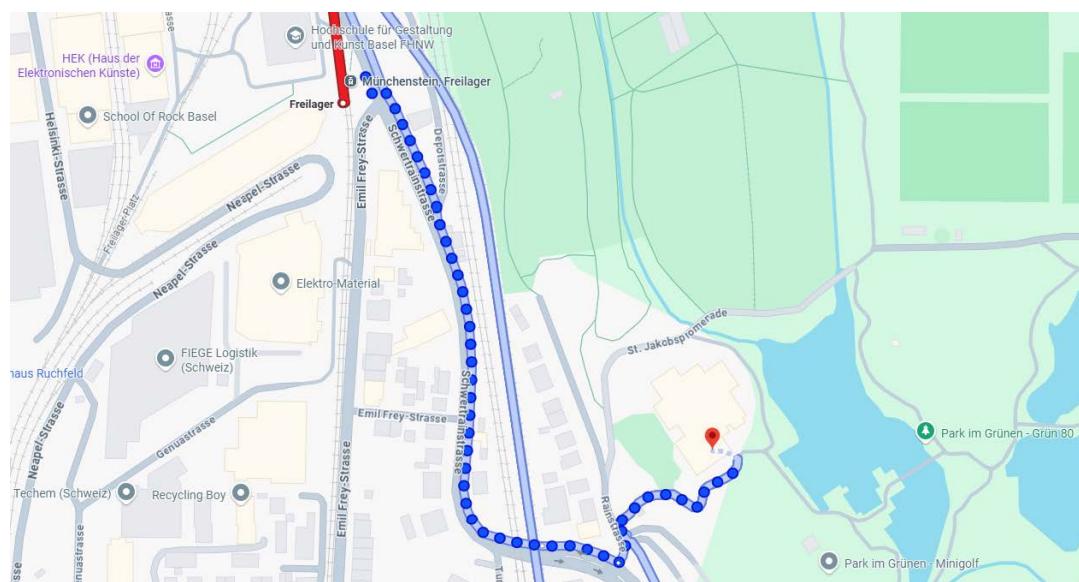
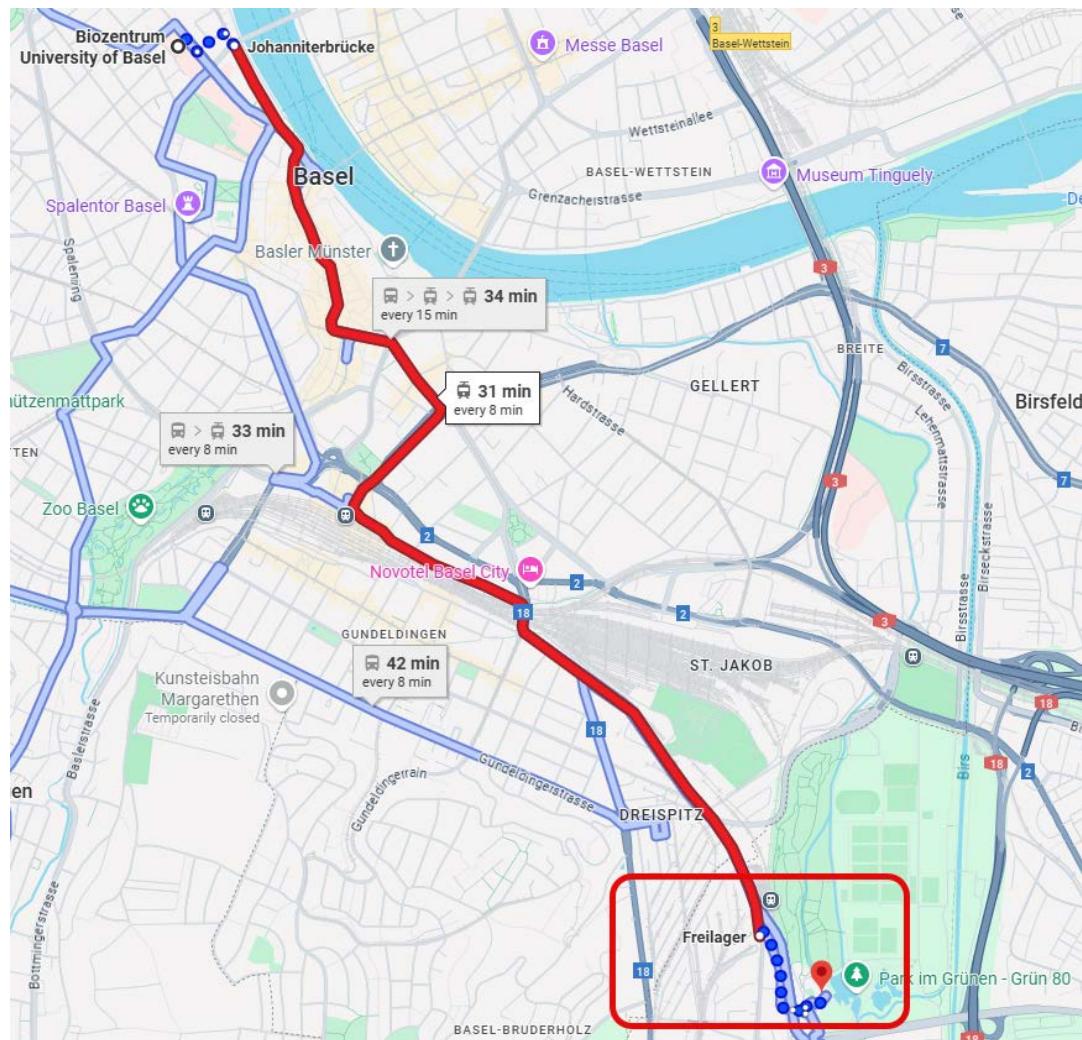


Restaurant Seegarten. Picture taken from [Park im Grünen](#).

How to get there by public transport:

- From Biozentrum University of Basel, Walk southeast on Spitalstrasse toward Wilhelm His-Strasse to reach *Johanniterbrücke* tram stop (approximately 4 minutes, 260 meters).
- Take Tram 11 in the direction of *Aesch BL, Dorf* and get off the tram at *Freilager* (approximately 21 minutes).
- Head south at *Freilager* following *Schwertrainstrasse* for approximately 5 minutes (500 meters). Turn left onto *Rainstrasse* to the entrance of the Conference dinner venue at Restaurant Seegarten.

General Information



Childcare

The conference venue at Biozentrum offers a breastfeeding room that can be used by participants during the conference. The key for the breastfeeding room can be picked up at the reception of Biozentrum.

There are no childcare facilities on site, however, there are external providers in Basel that can be contacted:

- [kindernaescht.ch](#) at Marktplatz (15 minutes from the conference venue at Biozentrum by foot) can host kids from 18 months to 12 years on an hourly to daily basis by prior arrangement.
- [familea.ch](#) at Freie Strasse (15 minutes from the conference venue at Biozentrum by foot): Call +41 61 260 82 82 to check availability. Once confirmed, [complete attached form](#). Please note: deadline for submitting form is 11th August 2025.

Biostats-tourism: Things to look out for

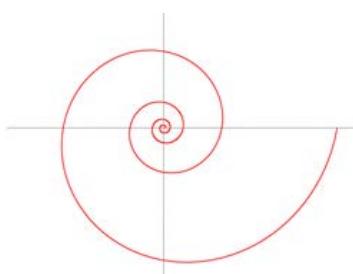
Jacob Bernoulli's tombstone

The cloister of the cathedral, Münsterplatz, Basel

Between around 1680 and 1800, eight members of the Bernoulli family were active in Basel and abroad as mathematicians and physicists. Jacob I (discovered the formula for the radius of curvature, the Bernoulli numbers, the fundamental theorem of probability theory, and the elastic curve etc) was so impressed by the equiangular or logarithmic spiral that he called it the *spira mirabilis*. He was impressed that the curve is "self-similar", that is any expansion of the curve will produce the same curve. He wrote that the spiral "may be used as a symbol, either of fortitude and constancy in adversity, or of the human body, which after all its changes, even after death, will be restored to its exact and perfect self". He wanted it carved on his tombstone with the Latin motto *Eadem mutata resurgo* (although changed, I will rise again the same). Unfortunately, the mason who did the carving misunderstood the instructions and gave him an Archimedean spiral instead.



Jacob Bernoulli (1686, around 32 years old), painted by his brother Nicolaus



Logarithmic spiral



Archimedean spiral in the epitaph of Jacob I (1654–1705)

House where Paracelsus lived

Totengässlein 3, now the Pharmacy Museum of the University of Basel

Paracelsus (real name Theophrastus von Hohenheim; 1493–1541) was a physician and alchemist. He is famous for revolutionizing medicine in the 16th century by proposing the use of one's own observations of nature rather than relying on ancient texts.





Leonhard Euler's childhood house

Kirchstrasse 7, Riehen, (from the outside only)

The mathematician Leonhard Euler (1707–1783) lived in the *Pfarrhaus* (clergy house) during his childhood. Euler was born in Basel, but his family moved to Riehen shortly after he was born. The house is close to the Fondation Beyeler art gallery.

Old University

Alte Universität, Rheinsprung 9, Basel (from the outside only)

Where the Bernoullis' and Paracelsus delivered lectures



Building where Daniel Bernoulli carried out experiments

Stachelschützenhaus (house of the crossbow men), Petersplatz, Basel (from the outside only)

Further buildings: Zur alten Treu, Nadelberg 15, home of Johann I Bernoulli
Englehof, Nadelberg 4, home of Johann II and Daniel Bernoulli
Imberhof, Andreasplatz 7-13, home of Nicolaus I Bernoulli

Sources: Bernoulli Euler Centre, University of Basel, Wikipedia, *Jakob Bernoulli's tomb*, MacTutor History of Mathematics Archive, available at: https://mathshistory.st-andrews.ac.uk/Extras/Bernoulli_tomb/. Speiser, David, "The Bernoullis in Basel," *Math Intelligencer*, 1992, Vol. 14, No. 4, pp. 46-47. PDF: <https://link.springer.com/content/pdf/10.1007/BF03024473.pdf>. Kitagawa, T. L., "The Origin of the Bernoulli Numbers: Mathematics in Basel and Edo in the Early Eighteenth Century," *Math Intelligencer*, 2022, Vol. 44, pp. 46–56. DOI: <https://doi.org/10.1007/s00283-021-10072-y>

Conference Schedule

Sunday 24 August 2025 | Pre-Conference Courses

	Biozentrum U1.111	Biozentrum U1.131	Biozentrum U1.141	Biozentrum U1.101	Biozentrum U1.197
09:00 - 10:30	Targeted Learning	The design of simulation studies	Network meta-analyses: from key concepts to advanced methods	Good Software Engineering Practice for R Packages	Bayesian Methods for Precision Medicine
10:30 - 11:00	Coffee break	Coffee break	Coffee break	Coffee break	Coffee break
11:00 - 12:30	Targeted Learning	The design of simulation studies	Network meta-analyses: from key concepts to advanced methods	Good Software Engineering Practice for R Packages	Bayesian Methods for Precision Medicine
12:30 - 13:30	Lunch	Lunch	Lunch	Lunch	Lunch
13:30 - 15:00	Targeted Learning	The design of simulation studies	Network meta-analyses: from key concepts to advanced methods	Good Software Engineering Practice for R Packages	Multi-state models: theory, applications, and new developments
15:00 - 15:30	Coffee break	Coffee break	Coffee break	Coffee break	Coffee break
15:30 - 17:00	Targeted Learning	The design of simulation studies	Network meta-analyses: from key concepts to advanced methods	Good Software Engineering Practice for R Packages	Multi-state models: theory, applications, and new developments

Monday 25 August 2025 | Main conference

	Biozentrum U1.111	Biozentrum U1.131	Biozentrum U1.141	Biozentrum U1.101	ETH E27	ETH E23	ETH E21
09:00 - 10:45	Formal opening Keynote: Confidence distributions	Livestream Opening and Keynote	Livestream Opening and Keynote	Livestream Opening and Keynote	Livestream Opening and Keynote	Livestream Opening and Keynote	Livestream Opening and Keynote
10:45 - 11:30	Coffee break	Coffee break	Coffee break	Coffee break	Coffee break	Coffee break	Coffee break
11:30 - 13:00	When worlds collide	Efficient analyses of clinical trials	Poster highlights	Dynamic borrowing and basket trials	Causal inference - dealing with bias	Survival analysis 1	Prediction / prognostic modelling 1
13:00 - 14:00	Lunch	Lunch	Lunch	Lunch	Lunch	Lunch	Lunch
14:00 - 15:30	AI-driven multi-omics data integration	Defining estimands for clinical trials	Adaptive and multi-arm multi-stage trials	Causal inference: target trial emulation	Meta-analyses 1	Prediction / prognostic modelling 2	Joint / longitudinal modelling
15:30 - 16:00	Coffee break	Coffee break	Coffee break	Coffee break	Coffee break	Coffee break	Coffee break
16:00 - 17:30	Prediction modelling meets causal inference	Estimands: causal and multiple imputation approaches	Clinical trials and regulatory issues	Model selection and simulations	Bayesian methods 1	Multi-omics data integration	Biomarker studies & diagnostic tests

Tuesday 26 August 2025 | Main conference

Wednesday 27 August 2025 | Main conference

	Biozentrum U1.111	Biozentrum U1.131	Biozentrum U1.141	Biozentrum U1.101	ETH E27	ETH E23	ETH E21
09:00 - 10:30	Cracking causal questions: Estimands for reliable and clinically relevant evidence	Design and evaluation of clinical trials	Competing events and multi-state modelling	Machine learning 2	Prediction / prognostic modelling 3	Infectious disease and longitudinal modelling	Biomarker studies & mixed topics
10:30 - 11:00	Coffee break	Coffee break	Coffee break	Coffee break	Coffee break	Coffee break	Coffee break
11:00 - 12:15	Plenary: Statistical & causal perspectives on ML for individualized treatment	Livestream plenary	Livestream plenary	Livestream plenary	Livestream plenary	Livestream plenary	Livestream plenary
12:15 - 13:30	ISCB Annual General Meeting						
13:00 - 14:00	Lunch	Lunch	Lunch	Lunch	Lunch	Lunch	Lunch
14:00 - 15:30	Improving replicability in clinical biostatistics	Randomization and analysis of clinical trials	Survival analysis 2	Causal inference: Mixed topics	Innovation in oncology dose escalation trials and beyond	Missing data and imputation	Observational/real-world data 1
15:30 - 16:00	Coffee break	Coffee break	Coffee break	Coffee break	Coffee break	Coffee break	Coffee break
16:00 - 17:30	Generative AI in clinical research and drug development (BBS invited session)	Efficient use of interim analyses in clinical trials	Survival analysis 3	Causal inference in time-varying settings	Prediction / prognostic modelling 4	Meta-analysis 3	Observational/real-world data 2

Conference Schedule

Thursday 28 August 2025 | Mini-Symposia

	Biozentrum U1.111	Biozentrum U1.131	ETH E21 & E 23
09:15 - 10:45	Causal Inference for Improved Clinical Collaborations: A practicum	STRATOS - Statistical Research needs to improve - on the important roles of simulation studies	Early Career Biostatisticians' (ECB) Day
10:45 - 11:30	Coffee break	Coffee break	Coffee break
11:30 - 13:00	Causal Inference for Improved Clinical Collaborations: A practicum	STRATOS - Statistical Research needs to improve - on the important roles of simulation studies	Early Career Biostatisticians' (ECB) Day
13:00 - 14:00	Lunch	Lunch	Lunch
14:00 - 15:30	Enhancing Cancer Clinical Trials with Patient-Reported Outcomes: Insights from SISAQOL-IMI		
15:30 - 16:00	Coffee break		
16:00 - 17:30	Enhancing Cancer Clinical Trials with Patient-Reported Outcomes: Insights from SISAQOL-IMI		

Pre-conference Courses

[CM] Half Day Course (morning): Bayesian Methods for Precision Medicine

Peter F. Thall

M.D. Anderson Cancer Center, United States of America

Time: Sunday, 24 August 09:00 - 12:30

Location: Biozentrum U1.197

This half day short course will present statistical concepts and methods related to precision, or ‘personalized’ medicine, which uses individual patient covariates to choose treatment or doses. The topics are drawn from the book, ‘Bayesian Precision Medicine’ published by Chapman and Hall in 2024. To start, the problem of comparing immunotherapy to prayer for treating a severe disease will be discussed. Basic concepts of causal inference will be reviewed, including bias correction methods for analyzing observational data, causal diagrams, with both toy and real-world illustrative examples. Two clinical trial designs will be reviewed that aim to identify optimal subgroup-specific doses or treatments in particular medical settings, each using a utility of a multivariate outcome. The first is a phase 1-2 design that uses the joint utility of five time-to-event outcomes to optimize patient subgroup-specific natural killer cell doses for treating advanced leukemia or lymphoma. The second design does phase 2 treatment screening and selection, illustrated by a randomized three-arm trial to compare targeted agents. Two data analyses that apply Bayesian nonparametric regression models to identify optimal covariate-specific treatments then will be presented. The first analysis uses observational data to identify optimal covariate-specific doses of intravenous busulfan as part of the preparative regimen for allogeneic stem cell transplantation. The second is a re-analysis of a published dataset from a randomized trial, with a joint utility of progression free survival time and total toxicity burden used to choose optimal personalized targeted therapies for advanced breast cancer

**[C1] Full Day Course 1:
Targeted Learning**

Oliver Dukes¹, Stijn Vansteelandt¹, Shaun Seaman²

¹Ghent University, Belgium

²University of Cambridge, United Kingdom

Time: Sunday, 24 August 09:00 - 17:00

Location: Biozentrum U1.111

Evaluating treatment effects using observational data or trials with complex intercurrent events may require accounting for high-dimensional confounding. This course will describe how, in these challenging situations, machine learning and variable selection procedures can be used to infer causal effects. The first part is a high level overview of how and why this methodology works, touching on recent developments in 'double machine learning' and 'targeted maximum likelihood estimation'. In the second part the participants will exercise the concepts and methods explained during the first part via the analysis of real data sets.

This introductory course is aimed at researchers in the (pharmaceutical) industry and academia working with observational as well as trial data; a basic understanding of causal inference can be helpful but is not necessary. We foresee a mix of lectures and hands-on exercises using R.

**[C2] Full Day Course 2:
The Design of Simulation Studies**

Tim P. Morris, Ian R. White

UCL, United Kingdom

Time: Sunday, 24 August 09:00 - 17:00

Location: Biozentrum U1.131

imulation studies are an invaluable tool for biostatistical research, as you will see at the ISCB46 conference this week, where they will play a prominent role in many presentations. You may even be presenting one of your own. Do you have the confidence to thoughtfully critique someone's simulation study, or to defend yours? Although it is tempting to think of a simulation studies as a mere coding exercise, we have a rather different view. This short course will focus on the design of simulation studies, following the ADEMP framework (Aims, Data-generating mechanisms, Estimands, Methods of Analysis, Performance measures). Practical sessions will be interactive rather than computer-based, involving group discussions and debates on issues around the design of simulation studies.

**[C3] Full Day Course 3:
Network Meta-Analysis: From Key Concepts to Advanced Methods**

**Virginia Chiocchia¹, Konstantina Chalkou¹, Orestis Efthimiou², Tasnim Hamza¹,
Georgia Salanti¹**

¹Institute of Social and Preventive Medicine, University of Bern, Switzerland

²Institute of Primary Health Care (BIHAM), University of Bern, Switzerland

Time: Sunday, 24 August 09:00 - 17:00

Location: Biozentrum U1.141

Network meta-analysis is an extension of pairwise meta-analysis that allows us to compare three or more interventions simultaneously, by combining direct and indirect evidence from a network of studies. Network meta-analysis can be used to estimate the relative treatment-effects between any pair of interventions in the network, it increases precision compared to using only direct evidence, and it can produce a hierarchy of the interventions.

In the morning session of this full-day course we will demonstrate the assumptions and methods of network meta-analysis and network meta-regression with interactive lectures and practical exercise. In the afternoon we will introduce advanced topics in network meta-analysis, such as the use of individual participant data, component network meta-analysis for composite interventions, and dose-response analysis.

The course is designed for participants who are familiar with meta-analysis and Bayesian statistics. By the end of the course, participants will be able to:

- Understand and assess the assumptions underlying the validity of indirect comparisons and network meta-analysis
- Estimate the relative treatment effects between any pair of interventions within a network of studies and present them in a transparent way
- Assess and test for inconsistency within a network of interventions
- Formulate a network meta-regression model and interpret the results and output
- Obtain an overview of advanced methods and extensions in network meta-analysis

The practical exercises will be performed in the statistical software R.

**[C4] Full Day Course 4:
Good Software Engineering Practice for R Packages**

Audrey Te-ying Yeo¹, Alessandro Gasparini², Daniel Sabanés Bové³

¹Finc Research

²Red Door Analytics AB

³RCONIS

Time: Sunday, 24 August 09:00 - 17:00

Location: Biozentrum U1.101

The vast majority of statisticians in academia and industry alike write statistical software daily. Nonetheless, software engineering principles are often neglected in biostatistics: most biostatisticians know a programming language (such as R) but lack formal training in writing reusable and reliable code.

This course aims to equip participants with the essential software engineering practices required to develop and maintain robust R packages. With the growing demand for reproducible research and the increasing complexity of statistical methods developed for multidimensional data, writing high-quality R packages has become a critical skill for statisticians to prototype, develop, and disseminate novel methods and push their adoption in practice. The course will focus on the key principles of software engineering, such as workflows, modular design, version control, testing, documentation, and quality indicators. Focussing on these aspects ensures the reliability and sustainability of R packages.

Participants will learn how to structure their R packages following best practices and making use of tools that streamline the development process. The course will also cover version control using Git, allowing participants to manage code changes effectively and collaborate with others. A significant emphasis will be placed on writing and running unit tests, ensuring that packages are error-free and behave as expected across different environments and over time.

Furthermore, the course will cover quality indicators for R packages and explore techniques for writing effective documentation, enabling users to pick, understand, and use statistical software packages effectively.

By the end of the course, participants will have a solid understanding of good software engineering principles tailored to R package development, enabling them to build packages that are not only functional but also reliable, reusable, and easy to maintain.

**[CA] Half Day Course (afternoon):
Multi-State Models: Theory, Applications and New Developments**

Liesbeth de Wreede, Hein Putter

Leiden University Medical Center, Netherlands

Time: Sunday, 24 August 13:30 - 17:00

Location: Biozentrum U1.197

Multi-state models play an increasingly important role in the analysis of time-to-event data. They provide a comprehensive framework to analyze and understand complex medical phenomena, making them invaluable in research aimed at improving patient care, guiding public health policies, and advancing medical science. Extensions of the Nelson-Aalen estimator of the cumulative hazard and of the Kaplan-Meier estimator of the survival function allow for a detailed assessment of the dynamics of complex disease processes and patient trajectories, and the effect of covariates on these patterns.

In the first half we offer a brief coverage of basic concepts and techniques in multi-state models, focusing on non- and semi-parametric (Cox-model based) Markov models. We start with an introduction to important concepts, in particular transition intensities (rates) and transition probabilities (risks) and the relation between them, viewing multi-state models as an extension of competing risks models. We continue with methods for assessment of the effect of covariates on the transition intensities through proportional hazards models. Throughout the course, all steps needed for a multi-state analysis will be illustrated with examples and syntax based on the mstate package in R.

In the second half we focus on two selected topics related to more recent advancements in multi-state modelling. The first concerns the Markov assumption. We discuss formal tests for the Markov assumption. We explore estimation of transition probabilities that are consistent also when the Markov assumption is violated, in particular the landmark Aalen-Johansen estimator, and extensions like the hybrid landmark Aalen-Johansen estimator. The second topic concerns incorporation of relative survival in multi-state models. This allows to split mortality in excess and background mortality with and without intermediate events. Two different models for assessing the impact of covariates on the excess hazard will be introduced: the Cox model and Aalen's additive hazards model.

Keynote Talks

Confidence distributions: new tools to design, adapt and analyse clinical trials

Monday, 2025-08-25 9:30 - 10:45, Biozentrum U1.111

Livestream of the plenary in all other rooms

Chair: Thomas Jaki (MRC Biostatistics Unit, University of Cambridge)

Ian Marschner, University of Sydney, Sydney, Australia

Ian Marschner is Professor of Biostatistics at the University of Sydney and Director of Biostatistics at the NHMRC Clinical Trials Centre in Sydney, Australia. He has over 30 years of experience as a biostatistician working on clinical trials research across many therapeutic areas, with a recent focus on innovative clinical trial design. He has published extensively on new statistical methodology for biostatistical applications and is a Chief Investigator for the Australian Trials Methodology Research Network. Formerly, he was Head of the Department of Statistics at Macquarie University, Director of Biometrics at Pfizer and Associate Professor of Biostatistics at Harvard University.



Confidence distributions provide a holistic summary of the information that the data contains about a parameter in a statistical model, expressed using a probability distribution over the parameter space. They provide a frequentist analogue of Bayesian posterior distributions, but without the requirement to specify a prior distribution. In randomised clinical trials, confidence distributions are particularly useful for summarising the evidence for a treatment effect, allowing the strength of evidence to be quantified using a confidence statement such as $\text{Conf}(\text{Benefit})=92\%$. In this talk, I will review the application of confidence distributions to clinical trials using various case studies and then present promising lines of future research. Confidence distributions are useful at all stages of a clinical trial, from design to monitoring to analysis. They are particularly promising for adaptive designs, where they can be used to adapt design features such as the randomisation probabilities, the sample size or the available treatments. These confidence-adaptive designs provide various advantages over other types of adaptive designs, by making use of connections with long-standing frequentist group sequential theory that allows less reliance on extensive simulation. In some contexts, confidence distributions may provide the advantages of a Bayesian analysis but with less complexity and sensitivity to assumptions. They have recently found their way into major medical journals and are a promising new tool for clinical biostatisticians.

Statistical and causal perspectives on machine learning in estimating individualized treatment strategies

Wednesday, 2025-08-27 11:00 - 12:15, Biozentrum U1.111

Livestream of the plenary in all other rooms

Chair: Vanessa Didelez (Leibniz Institute for Prevention, BIPS)

Erica E. M. Moodie, Biostatistics, McGill University, Canada

Erica E. M. Moodie is a Professor of Biostatistics and a Canada Research Chair (Tier 1) in Statistical Methods for Precision Medicine. She obtained her MPhil in Epidemiology in 2001 from the University of Cambridge and a PhD in Biostatistics in 2006 from the University of Washington, before joining the faculty at McGill. Her main research interests are in causal inference and longitudinal data with a focus on precision medicine. She is the 2020 recipient of the CRM-SSC Prize in Statistics and an Elected Member of the International Statistical Institute. Dr Moodie is a Co-Editor of *Biometrics*, a Statistical Editor of *Journal of Infectious Diseases*, and the 2024-2025 President of the Statistical Society of Canada.



The predictive power of machine learning is often celebrated, but caution is also warranted due to the potential for algorithmic bias which often arises from classical statistical concerns such as confounding and selection bias. Statisticians are thus often wary of the use of machine learning in the context of treatment recommendations and other highly sensitive and potentially life-altering decision-making.

I will discuss two examples in which machine learning approaches were incorporated into a classical statistical method to learn individualized treatment strategies that are designed to address confounding to yield causally valid conclusions. In the first, non-parametric ensemble learner is used in the context of an approach that is not robust to model mis-specification. In the second, probabilistic supervised learning in the form of Gaussian processes will be used to improve performance of inverse probability of treatment weighted estimators. In both cases, the use of machine learning is layered onto classical statistical approaches to causal inference that have been developed to address confounding in the context of observational data analysis. Relevant causal assumptions, and how they may (or may not!) be detected and mitigated will also be discussed.

Invited Sessions

When Worlds Collide: Common Methodological Themes in Meta-Analysis, Causal Inference, and Hybrid Trial Design

Monday, 2025-08-25 11:30 - 13:00, Biozentrum U1.111

Chair: David Phillipo

1: Combining Information from Multiple and Diverse Sources to Answer Causal Questions

Issa J. Dahabreh

Harvard University, United States of America

In this talk I will discuss approaches for structuring and conducting analyses that combine information from multiple and diverse sources to address questions about the effectiveness of medical interventions. By considering two practical examples — (1) extending inferences from a clinical trial to a “real-world” population and (2) augmenting a clinical trial with external data to improve efficiency in estimating treatment effects — I argue that combining information from multiple sources to answer causal questions requires novel study designs, causal assumptions, and statistical methods.

2: Doubly Robust Augmented Entropy Balancing for Externally Controlled Clinical Trials and Indirect Treatment Comparisons

Antonio Remiro-Azocar

Novo Nordisk, Spain

Background Conducting a properly-powered randomised controlled trial is not feasible in certain settings, due to small populations and a lack of trial-eligible patients, for life-threatening

and severely debilitating conditions with high unmet need, and for ethical reasons. Regulators recognize that externally controlled clinical trials may be required in special circumstances, and marketing authorisation applications featuring externally controlled trials are increasing. The reliance of payers and health technology assessment bodies on such research designs is also growing. For instance, in the absence of head-to-head comparisons between all relevant comparators, “unanchored” indirect comparisons are often required in therapeutic areas with a rapidly evolving treatment landscape.

Methods Various statistical methods have been proposed to adjust for imbalances in baseline covariates between the clinical trial and the external control. The most widely-used methodologies are singly robust propensity score-based weighting and outcome modelling-based approaches. Alternative weighting methods based on entropy balancing will be presented, which directly enforce covariate balance and are generally more stable, precise and robust to model misspecification than the standard “modelling” approaches to weighting. The presentation will introduce doubly robust estimators that augment the entropy balancing approaches by fitting a model for the conditional outcome expectation, then combining the predictions of the outcome model with the entropy balancing weights. The methods are evaluated in a simulation study and their application illustrated in an example analysis.

Results and Conclusions The presentation integrates parallel developments in the areas of indirect treatment comparisons (meta-analysis) and causal inference, under a unified framework for target estimands and covariate adjustment. Decision-makers have expressed a preference for doubly robust estimation approaches that can consistently estimate treatment effects as long as either a propensity score model or an outcome model is correct, but not necessarily both. Augmented entropy balancing-based estimators that are doubly robust and more bias-robust than commonly used approaches, as demonstrated by simulations involving binary outcomes, are presented.

3: Meta-Analytic Framework Using Individual Patient Data: A Case Study from Alopecia Arrieta

Satrajit Roychoudhury

Pfizer Inc., United States of America

The recent 21st Century Cures Act propagates innovations to accelerate the discovery, development, and delivery of 21st century cures. It includes the broader application of Bayesian statistics and the use of evidence from clinical expertise. An example of the latter is the use of trial-external (or historical) data, which promises more efficient or ethical trial designs.

Invited Sessions

Draft guidance FDA “Considerations for the Design and Conduct of Externally Controlled Trials for Drug and Biological Products” emphasized that sponsors must include patient-level data in the market-application. Though there are considerable literature discussing Bayesian methods to include summary-level external control data in design and analysis, the literature handling individual-level external control data is still small. We provided a robust meta-analytic framework when individual patient-level data is available. analyze individual patient data from various sources. We suggest a two-step approach. In the first step, the data from each source is analyzed in isolation, resulting in an estimate (and standard error) for the main parameter of the target population, taking account of the covariate information from each patient. In the second step, these adjusted data are then analyzed using robust Bayesian hierarchical model. The utility of the method is illustrated using simulation and a case study from Alopecia Areata area. The talk will further reflect on the regulatory discussions.

AI-Driven Multi-Omics Data Integration

Monday, 2025-08-25 14:00 - 15:30, Biozentrum U1.111

Chair: Niko Beerenwinkel

1: Generative AI for Unlocking the Complexity of Cells

Maria Brbic

EPFL, Switzerland

We are witnessing an AI revolution. At the heart of this revolution are generative AI models that, powered by advanced architectures and large datasets, are transforming AI across a variety of disciplines. But how can AI facilitate and eventually enable groundbreaking discoveries in life sciences? How can it bring us closer to understanding biology – the functions of our cells, their alterations in diseases, and variations across species? In this talk, I will show how generative AI can uncover spatial relationships between cells, enabling the reassembly of tissues from dissociated single cells. Next, I will discuss the future of discovery in the era of generative AI and foundation models, highlighting the paradigm shift in machine learning required to revolutionize biology.

2: Machine Learning for Multi-Omics Integration: Advancing Rare Disease Diagnostics in Pediatric Acute Care

Julia Vogt

Department of Computer Science, ETH Zurich, Switzerland

Rare diseases predominantly affect children, often causing premature death or lifelong disability. Over the past decade, the discovery of rare diseases has accelerated, driven by advances in genomic data generation and analysis, which now offer faster turnaround times at lower costs. This has led to a paradigm shift in the role of genomics in pediatric acute care. Despite these advancements, genome-dependent diagnostic rates remain low. As sequencing depth and speed have increased, the primary challenge has shifted from detecting genetic alterations to understanding their functional relevance. In this presentation, we will discuss novel deep learning methods to integrate multi-omics data—including whole-genome

sequencing (WGS), RNA sequencing (RNA-seq), proteotyping, and metabolomics—to identify rare, deleterious genetic variants in children with life-threatening extreme phenotypes. We will also demonstrate the clinical impact of these methods by analyzing their effects at the gene, transcript, and proteome/metabolome levels.

3: Correspondence Problems in Multi-Omic Data

Kjøng-Van Lehmann

Faculty of Mathematics, Computer Science and Natural Sciences, RWTH Aachen, Pauwelsstr 19, 52074 Aachen, Germany

The ability to generate multi-omic profiles at scale has created new opportunities to investigate the relationships across different omic read-outs. However, it comes with an increase in complexity making data analysis and interpretation more challenging. This heterogeneity of multimodal data pose significant analytical and interpretative challenges. Specifically, multi-modality creates correspondence problems, complicating the effective combination and analysis of diverse data types such as genomics, transcriptomics, proteomics data. Despite these challenges, leveraging synergies across datatypes may hold the potential to discover new insights by enabling a more comprehensive view of the underlying biological mechanisms. At single-cell resolution, the integration of multi-omics data allows us to leverage neural network architectures that can capture complex patterns and interactions within the data. In this talk, I will demonstrate examples of multi-omics integration and address correspondence problems demonstrating how we have leveraged the data to gain insights into biological mechanisms.

Prediction Modelling Meets Causal Inference for Clinical Decision Making

Monday, 2025-08-25 16:00 - 17:30, Biozentrum U1.111

Chair: Ruth Keogh

1: From Prediction to Treatment Decision: Aligning Development, Evaluation and Monitoring

Wouter A.C. van Amsterdam

University Medical Center Utrecht, Netherlands, The

From sepsis prediction to heart-attack risk and cancer prognosis, the medical literature is full of models predicting future patient health. Many of these models are motivated by the goal of supporting clinical decisions, yet few are designed or evaluated with this goal in mind.

Prediction models are typically assessed based on their predictive accuracy, but strong predictive performance does not automatically translate into better clinical decision-making. To truly optimize patient outcomes, we must align model development, evaluation, and monitoring with the goal of informing treatment decisions.

In this talk, I will outline a framework for this alignment, discussing key pitfalls in traditional predictive model evaluation and how to assess real-world decision impact. I will also address the challenge of monitoring predictive models in clinical practice—an area where EMA and FDA regulations demand action but offer little guidance. Standard metrics like discrimination and calibration can be misleading when applied to deployed models, particularly when not recognizing the effect of the deployed model on treatment decisions and patient outcomes.

Finally, I will discuss prediction-under-intervention models—models designed to estimate patient outcomes under different treatment scenarios. These models directly connect predictions to treatment decisions, making evaluation and monitoring conceptually straightforward. Though conceptually appealing, these models come with their own challenges—both methodological and practical.

Mapping out approaches to predictive model evaluation and monitoring—and addressing the lack of clear guidance—will be essential to ensuring that predictive models truly support clinical decision-making.

2: Dynamic Prediction of Survival Benefit to Inform Liver Transplant Decisions in Hepatocellular Carcinoma Patients

Pedro Miranda Afonso¹, Hau Liu², Michele Molinari³, Dimitris Rizopoulos¹

¹Department of Biostatistics, Erasmus MC, The Netherlands

²Starzl Transplant Institute, University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania, US

³J.C. Walter Jr Transplant Center, Houston Methodist Hospital, Houston, Texas, US

Background/Introduction Liver transplantation (LT) is the only curative treatment for selected patients with unresectable hepatocellular carcinoma (HCC). However, due to organ scarcity, patients must often wait for a suitable graft, during which they may become ineligible due to tumour progression or clinical deterioration. A predictive model identifying patients at the highest risk of waitlist dropout and those who would benefit most from LT could improve organ allocation.

Transplant-related survival benefit, defined as the additional survival time gained from LT compared to waitlist survival, provides a comprehensive metric to guide allocation. Estimating this causal effect requires addressing the observational nature of transplant data and time-varying confounders. To address these challenges, we developed a joint model for longitudinal and time-to-event data that dynamically predicts individualised transplant-related survival benefit in HCC patients. Unlike alternative approaches, such as the G-formula, structural marginal models and targeted maximum likelihood estimation, our model makes stronger assumptions about the biomarker measurement process but remains non-parametric for competing processes like censoring and visit times.

Methods We analysed data from 7,471 HCC patients listed in the US Scientific Registry for Transplant Recipients (SRTR) between 2012 and 2022, of whom 4,786 received a liver. We developed a Bayesian joint model to associate the pre-transplant trajectories of three well established predictors—the serum level of tumour α -fetoprotein (AFP) level, the tumour burden score (TBS), and the model for end-stage liver disease (MELD) score—with the risk of death before and after transplantation. We defined the assumptions necessary to obtain unbiased estimates of the causal effect of transplantation using observational data. Our model predicts a patient's survival probabilities with and without transplantation, which are then used to estimate liver transplant survival benefit. Dynamic updates enable real-time refinement of the predictions and identification of the patients most likely to benefit from transplantation. The model is implemented in the freely available R package JMbayes2.

Results Our model reveals distinct forms of association between AFP, TBS, and MELD score and the risk of death before and after transplantation. It provides unbiased estimates

of the causal effect of transplantation on individual survival using observational SRTR data without explicitly requiring a model for the transplant assignment mechanism.

Conclusion This prediction model represents an advancement in optimizing liver transplant decisions, promoting fairer organ allocation, and improving overall survival for waitlisted HCC patients.

3: Clinical Applications of Predictions under Interventions

Nan van Geloven

Leiden University Medical Center, Leiden, The Netherlands

Prognostic models (or prognostic algorithms) are increasingly used to inform medical treatment decisions. Typically, individuals with high risks of adverse outcomes are advised to start (or intensify) treatment while those at low risk are advised more conservative treatment. Regular prediction models do not always provide risks that are relevant to inform such decisions: for example, an individual may be estimated to be at low risk because similar individuals in the past received a treatment which lowered their risk. To overcome these limitations, new proposals focus on predicting outcomes under specified treatment options. These are known as counterfactual predictions or predictions under interventions. Estimating and evaluating predictions under interventions using observational data comes with additional requirements such as causal assumptions, confounding adjustment, as well as suitable data.

In this talk, I will illustrate several clinical applications where prediction under interventions were used, and contrast them to the information one could get from regular prediction models. I will start with a simple point treatment setting where confounding variables are already part of the predictor set and explain why even here regular ways of predictive model development and evaluation may not be sufficient. Following applications will build up in complexity, including settings with time-varying treatments, time-varying confounding as well as sequential (i.e. repeated) decision making. For each application, I will focus on why predictions are needed under certain (or multiple) intervention strategies and point out the additional data requirements and steps needed during analysis. Clinical applications include fertility, cardiology, transplantation and transfusion medicine.

Mathematical and Statistical Modelling in the Life Sciences: Seeking Causal Explanations

Tuesday, 2025-08-26 09:15 - 10:45, Biozentrum U1.111

Chair: Jack Kuipers

1: In Silico Simulations of Cancer-Immune Interactions to Aid Clinical Trial Design and Execution

Johannes Textor, Jeroen Creemers

Radboud University Nijmegen, The Netherlands

Cancer immunotherapy is an important application area for mathematical modeling. Current modeling studies have a range of ambitious goals from dose optimization to creating “digital twins” of individual cancer patients for treatment response prediction. Here we focus on a humbler, but nonetheless important, goal: aiding with the planning and design of clinical trials. Cancer immunotherapy trials can be hard to design due to heterogeneous and time-varying treatment effects. While clinical statisticians already use computer simulations, these rarely integrate explicit pathophysiological mechanisms, such as cancer-immune interactions, to specifically adapt the design to the treatment. Encouraged by rapid progress in mathematical modeling, we here propose an “in-silico-first” approach—already common in industry—where doctors, statisticians, and modelers build knowledge-based mathematical models to examine and refine the statistical design of clinical trials. We will discuss some experiences with such in silico trial designs in the cancer immunotherapy field. We will also, reflect on how these approaches contrast to using causal diagrams for similar purposes, and offer some opinions on what we feel are advantages and disadvantages of each approach.

2: Evolutionary Dynamics of Cancer Progression and Response to Treatment

Ivana Bozic

University of Washington, United States of America

Cancer results from a stochastic evolutionary process characterized by the accumulation of mutations that are responsible for tumor initiation, progression, immune escape, and drug

resistance, as well as mutations with no effect on the phenotype. Mathematical modeling, combined with clinical, sequencing and epidemiological data, can be used to describe the dynamics of tumor cell populations and to obtain insights into the hidden evolutionary processes leading to cancer. I will present recent approaches for quantifying the evolutionary dynamics of cancer in patients, and their implications for deciphering cancer heterogeneity and response to therapy.

3: Using DAGs and Dynamical Simulations to Explore and Understand Causal Links

Theis Lange

Department of Public Health, University of Copenhagen, Denmark

Directed acyclic graphs (DAGs) play a large role in the modern approach to causal inference. DAGs describe the relationship between measurements taken at various discrete times including the effect of interventions. The causal mechanisms, on the other hand, would naturally be assumed to be a continuous process operating over time in a cause–effect fashion. At the same time DAGs are less suited to provide insight into the actual effects of a given intervention on a desired outcome. The reason being that DAGs are really just an encoding of direction and absence of causal effect. It does not naturally express magnitudes or even directions. Finally, there are multiple misconceptions on if/how feedback can be expressed in DAGs. In this talk I will present how classic DAGs can be combined with systems dynamic thinking and simulations to increase their usability. It will also explore how the underlying principles in causal inference thinking (eg keep the scientific question in focus as well as clearly defined) can also strengthen simulation based methos.

Optimizing Efficiency in Adaptive Trial Designs

Tuesday, 2025-08-26 11:30 - 13:00, Biozentrum U1.111

Chair: Katherine Lee

1: How can We Foster a Sensible Use of Adaptive Designs for Regulatory Decision-Making?

Frank Petavy

EMA

Since 2007 there are recommendations for EU regulatory assessors and sponsors seeking marketing approval on trials with design modifications based on the results of an interim analysis in the Reflection Paper on Methodological Issues in Confirmatory Clinical Trials Planned with an Adaptive Design. In 2019 a harmonisation process was initiated between worldwide regulatory agencies, Industry associations and other public health stakeholders. This ICH guideline focuses on the design, conduct, analysis, and interpretation of adaptive clinical trials and is expected to provide a transparent and harmonised set of principles for the regulatory review of these studies in a global drug development programme. This presentation will explain the opportunity given by ICH E20 to rediscuss principles for adaptations in confirmatory clinical trials and describe the key considerations for these types of clinical trial in a regulatory setting, in order to motivate comments on the draft ICH E20 during the public consultation.

2: Optimizing Treatment Allocation in Platform Trials: A Need for New Rules?

Marta Bofill Roig¹, Martin Posch²

¹Universitat Politècnica de Catalunya, Spain

²Medical University of Vienna, Austria

Platform trials are randomized clinical trials designed to simultaneously compare multiple interventions, typically against a common control group. Arms to test experimental interventions may enter and leave the platform over time. Therefore, the number of experimental intervention arms in the trial can change over time. Determining the optimal allocation rates

for assigning patients to the treatment and control arms in platform trials poses a challenge. As the treatment arms enter or exit the platform, the optimal allocation rates also need to be adjusted. Additionally, the optimal allocation strategy depends on the specific analysis method used.

In this talk, we describe optimal treatment allocation rates for platform trials with shared controls, assuming that a stratified estimation and testing procedure based on a regression model is used to adjust for time trends. We consider analysis methods using concurrent controls only as well as methods based on also non-concurrent controls. We show that to minimize the maximum of the variances of the effect estimators, the optimal solution depends on the entry time of the arms in the trial. Generally, this solution does not correspond to the square root of k allocation rule used in the classical multi-arm trials. We illustrate the optimal allocation and evaluate the power and type 1 error rate compared to trials using one-to-one and square root of k allocations. In addition, we will discuss extensions, such as allocations in trials with varying eligibility and inclusion criteria and optimal allocation to the control group, for designs where the allocation rates to the treatment arms are equal.

3: Confirmatory Adaptive Enrichment Designs with a Continuous Biomarker

Nigel Stallard

University of Warwick, UK, United Kingdom

Background With the growing importance of clinical trials in targeted medicine there has been recent interest in adaptive enrichment designs [1]. In these two-stage designs patients from the first stage are used to identify a biomarker-defined population in which a treatment effect is anticipated. In the second stage the trial population is ‘enriched’ by restricting recruitment to patients from the selected population. At the end of the trial a hypothesis test is conducted of the treatment effect in the selected population. The data-dependent selection leads to statistical challenges if data from both stages are used for this hypothesis test.

Methods If the biomarker is measured on a continuous scale, with any biomarker by treatment interaction assumed to be monotonic, population selection is equivalently to identification of a cut-point for the biomarker. In this case subgroups considered are nested and the multiple testing procedure can be considered in a hierarchical fashion, enabling control of the familywise type I error rate (FWER) through a simple closed testing procedure given a valid test based on data from each possible selected subgroup.

Invited Sessions

Focussing on the case in which the outcome is normally distributed, two methods for this test are proposed. The first assumes selection of the subgroup with the largest test statistic when the distribution of this test statistic can be obtained by considering the joint distribution of the test statistics from different subgroups [2]. In the second selection is based on a fitted linear relationship between the outcome and the continuous biomarker [3].

Results The second approach proposed is shown to be more powerful when the linear model assumptions are met, but can lead to type I error rate inflation when they are violated whereas the former method can be less powerful but provides FWER control irrespective of the relationship between the biomarker and response [3].

References [1] Simon, N., Simon, R. Adaptive enrichment designs for clinical trials. *Biostatistics*, 14, 2013, 613-625.

[2] Stallard, N. Adaptive enrichment designs with a continuous biomarker. *Biometrics*, 79, 2023, 9-19.

[3] Stallard, N. Testing for a treatment effect in a selected subgroup. *Statistical Methods in Medical Research*, 33, 2024, 1967-1978.

Cracking Causal Questions: Estimands for Reliable and Clinically Relevant Evidence

Wednesday, 2025-08-27 09:00 - 10:30, Biozentrum U1.111

Chair: Marcel Wolbers

1: Chasing Shadows: How Implausible Assumptions Skew Our Understanding of Causal Estimands

Kelly Van Lancker, Stijn Vansteelandt

Ghent University, Belgium

The ICH E9 (R1) addendum on estimands, coupled with recent advancements in causal inference, has prompted a shift towards using model-free treatment effect estimands that are more closely aligned with the underlying scientific question. This represents a departure from traditional, model-dependent approaches where the statistical model often overshadows the inquiry itself. While this shift is a positive development, it has unintentionally led to the prioritization of an estimand's ability to perfectly answer the key scientific question over its practical learnability from data under plausible assumptions. We illustrate this by scrutinizing assumptions in the recent clinical trials literature on principal stratum estimands, demonstrating that some popular assumptions are not only implausible but often inevitably violated. We advocate for a more balanced approach to estimand formulation, one that carefully considers both the scientific relevance and the practical feasibility of estimation under realistic conditions.

2: Paving the Ground for Breakthrough Innovations in Alzheimer's Disease Drug Development with the Estimand Framework and Causal Inference

Paul Delmar, Marion Monchalin, Rachid Abbas

F.Hoffmann-La Roche Ltd., Switzerland

The adoption of the estimand framework within the clinical trial community, extending beyond data science, has made significant progress. The extensive educational efforts to generalize the understanding and application of causal concepts and the estimand framework are now

facilitating more effective and meaningful cross-functional conversations. In healthcare companies, these efforts also guide the allocation of limited resources dedicated to biostatistical innovations towards the most impactful.

Drawing on experience from early and late-stage clinical development in Alzheimer's disease and other neurological conditions, this presentation will review real-life examples of how this paradigm shift has impacted the design, conduct, analysis and interpretation of clinical trials. We will look into future areas of development and opportunities for advancing the field.

3: Estimating Causal Overall Survival Estimands in the Presence of Treatment Switching using Multi-State Models

Alexandra Bühler², Tobias Mütze¹, Lisa Hampson¹, Jiali Wang¹, Tomas Haas¹

¹Novartis Pharma AG

²University of Waterloo

In phase III oncology trials, the analysis of overall survival (OS) is often complicated by unidirectional treatment crossover, where patients randomized to control are permitted to switch to the experimental treatment upon disease progression. While a treatment policy estimand is routinely adopted as the primary analysis in this setting, hypothetical estimands have gained considerable attention as supplemental analyses; most relevant are estimands that contrast the "*experimental treatment*" with the hypothetical regime of "*control treatment without the option to crossover after progression*". Rank preserving structural failure time models (RPSFTM) and inverse probability weighting (IPW) methods remain popular for estimating such estimands, but they all rely on unverifiable assumptions and have their limitations.

In this talk, we illustrate the utility of multi-state models for constructing and estimating marginal, causally interpretable OS estimands in the hypothetical scenario that crossover is not permitted. Specifically, following the idea of Gran et al. (2015), we define hypothetical treatment regimes by artificially manipulating transition intensities in the observed multi-state process and estimating OS probabilities by g-computation under some common causal assumptions. We report simulation results to demonstrate the performance of the proposed approach in realistic clinical settings and present a roadmap for its implementation in R. We discuss potential extensions to other relevant hypothetical treatment regimes. The talk will conclude with remarks on the robustness of modeling assumptions required by the method.

Improving Replicability in Clinical Biostatistics

Wednesday, 2025-08-27 14:00 - 15:30, Biozentrum U1.111

Chair: Anne-Laure Boulesteix

1: Questionable Research Practices - From Small Errors to Research Misconduct

Leonhard Held

University of Zurich, Switzerland

The pressure to 'publish or perish' increases the chances that researchers report results selectively, apply data dredging, or even try to cheat the system. It is helpful to consider such Questionable Research Practices (QRPs) as a spectrum of behaviours, ranging from honest errors and mistakes at one end, through to misconduct and fraud at the other. I will give some recent examples of the spectrum of QRPs from the biomedical literature. As the number of research paper retractions currently on the rise, we can no longer dismiss QRPs as isolated problems of a small number of people behaving sloppily or dishonestly. Instead, every researcher and every statistician may at times engage in QRPs and hence should be aware of the various forms in their and others' research. Addressing QRPs should be a central part of our identity as biostatisticians to facilitate rigorous, transparent, and reproducible research practices.

2: Improving the Replicability of Applied and Methodological Research

Sabine Hoffmann

Ludwig-Maximilians-University Munich, Germany

In recent years, there has been increasing awareness that result-dependent selective reporting among a multiplicity of possible analysis strategies leads to unreplicable research findings. While the use of better statistical methods is one of the solutions that is often suggested to improve the replicability of research findings, there is also evidence that methodological research is itself not immune to incentives encouraging result-dependent selective reporting. This talk will introduce different topics that are relevant for the replicability of research findings in applied and methodological research and give an overview of potential solutions to

improve the replicability of research findings, ranging from pre-registration, registered reports and blind analysis to multiverse style analyses, multi-analyst studies and neutral simulation studies.

3: Method Benchmarking in Computational Biology - Current State and Future Perspectives

Charlotte Soneson

Friedrich Miescher Institute for Biomedical Research (Switzerland), SIB Swiss Institute of Bioinformatics (Switzerland)

Researchers, regardless of discipline, are often faced with a choice between multiple computational methods when performing data analyses. Method benchmarking aims to rigorously compare the performance of different methods, typically using ground truth derived from well-characterized reference datasets, in order to determine the strengths and weaknesses of each method or to provide recommendations regarding suitable choices of methods for a specific analysis task. In this talk I will discuss the current state of benchmarking in computational biology, using examples from the field of single-cell data analysis. I will discuss challenges, as well as ideas for making benchmarking more reproducible, extensible and continuous.

4: Software Sensitivity Analysis in Medical and Methodological Research

Tim P. Morris

UCL, United Kingdom

Principled sensitivity analysis helps researchers assess the sensitivity of our inference to assumptions. Assumptions which spring to mind may be normality or independence, which are inherent to a particular statistical method. However, a software implementation may make further ‘assumptions’ through its default settings. High quality statistical software gives users control over options/arguments but makes defensible default choices otherwise. It is not unusual to have more than one defensible choice, making defaults somewhat arbitrary. The collection of default choices used by different software implementations may in aggregate lead to software sensitivity for a given method. I will present some collected examples of

Invited Sessions

such ‘software sensitivity’ in medical and methodological research, and argue – to myself as much as others – for us to consider it more routinely.

Generative AI in Clinical Research and Drug Development (BBS Invited Session)

Wednesday, 2025-08-27 16:00 - 17:30, Biozentrum U1.111

Chair: Giusi Moffa

1: (Generative) AI in Medicine: Where Does Statistics Come into Play?

Sarah Friedrich

University of Augsburg, Germany

Artificial intelligence, including generative models, is transforming medicine through applications in diagnostics, treatment planning, and research. However, the success of these technologies relies not only on advances in machine learning but also on robust statistical foundations. From study design and data quality assessment to the differentiation between correlation and causation, statistics ensures the reliability and interpretability of AI-driven decisions in clinical practice. One key aspect is the evaluation and comparison of AI methods – and the data used for this. While supervised learning typically relies on benchmarking datasets, statistical approaches usually focus on simulation studies. Each approach has its strengths and limitations: benchmarking ensures comparability but may lack generalisability, whereas simulations allow controlled experimentation but may not fully capture real-world complexity. A related aspect is the growing field of synthetic data. It comes with the promise of preserving privacy and expanding the size of training samples in scenarios where obtaining real data is costly, challenging, or infeasible. However, ensuring that it faithfully represents real-world distributions is not straightforward.

In this talk, we will discuss the role statistics plays in AI, particularly focusing on the role of the data basis for training and comparison of models. We will review the advantages and disadvantages of benchmarking vs. simulation studies and touch upon properties, promises and challenges of synthetic data.

2: Do Single-Cell Transformers Understand Gene Regulation? a Cautionary Benchmark

Simon Anders¹, Constantin Ahlmann-Eltze¹, Wolfgang Huber²

¹University of Heidelberg, Germany

²EMBL Heidelberg, Germany

Deep networks are used with great success in some areas of cell biology, such as image-analysis tasks, but progress for functional genomics remains at best mixed. Yet, several recent studies claimed that transformer models, pre-trained on millions of single-cell transcriptomes, can predict the transcriptional outcome of CRISPR Perturb-Seq screens. If true, this would not only allow for replacing wet-lab experiments with in-silico reasoning on a large scale, but also imply that these models have gained actual understanding of the dynamics of gene regulation.

We have benchmarked these recently published pre-trained models against a deliberatively naive linear model that by design cannot capture gene-gene interactions. We found that the deep networks failed to outperform this linear base line. Therefore, claims that deep models can predict unseen experiments must be considered premature.

I will present our study and discuss possible reasons why the models fell short: for example, one key factor might be that transcriptomes were pooled across experiments and provided without metadata. This raises the question what network architecture we would need so allow feeding a model information about the hierarchical structure of the data and about experimental covariates. I will also compare with models that work with DNA sequence data, as there, much progress has recently been achieved in predicting, with cell-type and -state specificity, epigenetic and chromatin features and their effect on gene expression. I will discuss how this highlights inherent differences between omics data and text, which prohibits a direct transfer of the transformer architecture from the (tremendously successful) large-language models to functional genomics tasks, and how progress will need to address these differences in data type by innovative new model architectures. Finally, we also need consensus on robust benchmark schemes, possibly with community challenges, in order to distinguish real progress from overfitting.

3: Panel Discussion

Panelists: Simon Anders, Lilla Di Scala, Sarah Friedrich, Holger Fröhlich (virtual), Kostas Sechidis, Mark M.A. van de Wiel, Maarten van Smeden (virtual)

Abstracts of Contributed Talks

Efficient Analyses of Clinical Trials

Monday, 2025-08-25 11:30 - 13:00, Biozentrum U1.131

Chair: Tomasz Burzykowski

1: The Power of Three. a Framework to Guide Analysis of Covariance in Randomised Clinical Trials.

Stephen John Senn^{1,2,3}, Franz Koenig², Martin Posch²

¹University of Sheffield, United Kingdom

²Medical University of Vienna

³University of St Andrews

It is generally recognised that incorporating prognostic factors in the model used to analyse a clinical trial can improve the efficiency of estimates. Exactly what the improvement might be has generally either been investigated asymptotically or by simulation. Here we show that by assuming that covariates are Normally distributed, it is possible to obtain exact theoretical results for any sample size for the three factors that govern efficiency, namely 1) the expected mean square error 2) the variance inflation factor (VIF) & 3) second order precision, that is to say the precision of the estimate of the mean square error.

Fairly obviously, the main influence on the expected mean square error is the partial correlation with outcome of any prognostic factor to be added to the model. Also, obviously second order precision is simply determined by the residual degrees of freedom. However, what influences the VIF is less obvious and we show that it is equal to $1-R_Z^2$, where R_Z^2 is the coefficient of determination for the treatment indicator, Z, using the covariates as predictors. We give an exact expression for the expected value of the VIF that depends simply on the sample size and the number of predictors. This should greatly assist trialists in planning their analyses.

We also give reasons for believing that these formulae will work well, even for gross violation of Normality by the covariates (for example, categorical covariates) and show a relation-

ship between the VIF and the treatment by category chi-square statistic for any categorical covariate. We suggest that this theoretical framework can 1) provide a means of guiding and interpreting simulation studies 2) shed light on many practical matters, for example whether it is worth adding a covariate to a model, what the value might be of using so-called super-covariates and whether stratifying continuous covariates is sensible.

These results should be of particular interest for trialists working in rare diseases, where patient numbers in trial are low and relying on asymptotic results can be misleading.

2: Stirring the Pot: Combining Influence Functions and Wald Type Tests for more Powerful Closed Testing Procedures

Christian Bressen Pipper, Klaus Holst

Novo Nordisk A/S, Denmark

We use influence functions of estimators to derive the large sample properties of a Wald type test for the intersection of two superiority hypotheses. This is done via the so-called stacking approach without making any assumptions on the simultaneous behavior of estimators. The resulting test is shown to have good power properties and thus forms the basis of a powerful closed testing procedure for testing two superiority hypotheses. We compare the proposal to the Bonferroni-Holm procedure and identify a number of scenarios in which superior performance is ensured. The actual power gain is investigated through simulations. Finally, we present an available software implementation through the R-package targeted and discuss how the methods can be extended to more than two superiority hypotheses.

3: Analyzing Multi-Center Randomized Trials with Covariate Adjustment While Accounting for Clustering

Muluneh Alene Addis, Kelly Van Lancker, Stijn Vansteelandt

Ghent University, Belgium

Augmented inverse probability weighting (AIPW) and G-computation with canonical generalized linear models have become increasingly popular for estimating the average treatment effect (ATE) in randomized experiments. These estimators leverage outcome prediction

models to adjust for imbalances in baseline covariates across treatment arms, improving statistical power compared to unadjusted analyses, while maintaining control over Type I error rates, even when the models are misspecified. Practical application of such estimators often overlooks the clustering present in multi-center clinical trials. Even when prediction models account for center effects, this omission can degrade the coverage of confidence intervals, reduce the efficiency of the estimators, and complicate the interpretation of the corresponding estimands. These issues are particularly pronounced for estimators of counterfactual means, though somewhat less severe for those of the ATE, as demonstrated through Monte Carlo simulations and supported by theoretical insights. To address these challenges, we develop efficient estimators of counterfactual means and of the ATE in a random center. These extract information from baseline covariates by relying on outcome prediction models, but remain unbiased in large samples when these models are misspecified. We also introduce an accompanying inference framework inspired by random-effects meta-analysis. Adjusting for center effects yields substantial gains in efficiency, especially when treatment effect heterogeneity across centers is large. Monte Carlo simulations and application to the WASH Benefits Bangladesh study demonstrate adequate performance of the proposed methods.

4: Increasing Efficiency of Composite Endpoint Trials: Novel Bayesian Latent Variable Framework with Application to Late-Stage Trials

Paul Newcombe¹, Jasna Cotic¹, Aris Perperoglou¹, James Wason², Dave Lunn¹

¹GSK, United Kingdom

²Newcastle University, United Kingdom

Composite responder endpoints, which combine multiple clinical outcomes to determine a binary responder variable, are commonly used in clinical trials to capture various aspects of disease progression. Traditionally these endpoints are analysed as binary, which means a large amount of information is discarded as the continuous component variables are dichotomised and collapsed together. Various methods, including a latent variable framework proposed by McMenamin et al[1], enable more efficient analysis of composite endpoints through an expanded model that includes the underlying continuous endpoint information to improve precision, while inferring treatment effects on the same composite endpoint scale. Previous applications to academic trials, and post-hoc analysis of pharmaceutical trial data, have indicated up to 60% reductions in sample size can be possible[1].

Despite clear potential to enable smaller, shorter trials, thereby decreasing costs and delivering new medicines to patients faster, there are no examples to our knowledge of this methodology being put forward with regulators, or being used to design a clinical trial within

the pharmaceutical industry. Implementing a Bayesian approach could increase uptake by enabling integration into quantitative decision-making frameworks such as conditional assurance[2], which are increasingly used during earlier phases of drug development. We will describe a novel Bayesian implementation of the composite responder latent variable framework, and demonstrate that it enables power increases as much as 50% in realistic simulations based on GSK trial data. Results illustrating application of both Bayesian and Frequentist implementations to secondary analysis of several GSK phase 3 trials will also be presented, which represents the first application of the latent variable framework to large, late stage trials and indicates a substantial sample size saving could have been possible. We hope to raise awareness of this important technique with statisticians working in all stages of drug development, and prompt further methodological development.

[1] McMenamin M, Barrett JK, Berglind A, Wason JM. Employing a latent variable framework to improve efficiency in composite endpoint analysis. *Stat Methods Med Res.* 2021. 30(3):702–16.

[2] Temple JR, Robertson JR. Conditional assurance: the answer to the questions that should be asked within drug development. *Pharm Stat.* 2021. 30(6): 1102-1111

5: Ordinal Outcome Analysis in Neurological Trials: Current Practices and the Proportional Odds Debate

Yongxi Long¹, Bart Jacobs², Ewout Steyerberg³, Erik van Zwet¹

¹Leiden University Medical Center, The Netherlands

²Erasmus Medical Center, The Netherlands

³University Medical Center Utrecht, The Netherlands

Ordinal scales, such as the modified Rankin Scale and Glasgow Outcome Scale Extended, are widely used as outcome measures in neurological trials. In a literature review of 70 recent randomized controlled trials (RCTs) across five acute neurological conditions, we examined statistical methods used to test and estimate treatment effects from ordinal outcomes.

Dichotomization remained common in about one-third of the RCTs, with notable discrepancy in the cut-point chosen for analysis. Therefore, the information contained in the rank ordering of the outcome was not fully used. Among studies that retained the ordinal nature of the data, the proportional odds model was commonly used to quantify the treatment effect in terms of a common odds ratio and to test the null hypothesis that the treatment has no effect. However, there is a large variation in the assessment and reporting of the proportional

Abstracts of Contributed Talks

odds assumption. This lack of clarity can lead to misinterpretation of results and suboptimal methodological choices. Concern over the validity of this assumption may lead researchers to dichotomize ordinal outcomes unnecessarily.

To address these challenges, we developed methodological guidance for ordinal outcome analysis, with a particular focus on the proportional odds assumption. We explain why the proportional odds assumption is irrelevant for hypothesis testing but crucial for summarizing treatment effects. We also demonstrate that pre-testing the proportional odds assumption can lead to inflated type I error rates. We advocate for a simple graphical check of the assumption as more informative than formal testing. If we are satisfied that there is no substantial violation of the proportional odds assumption, it is reasonable to summarize the treatment effect into a single number such as the common odds ratio.

We illustrate these considerations with three neurological trials: the ANGEL-ASPECT and the MR CLEAN trial (which investigated endovascular therapy in stroke) and the RESCUEicp trial (which investigated decompressive craniectomy in traumatic brain injury). We conclude with a statistical checklist for ordinal outcome analysis. By addressing common misconceptions and providing practical recommendations, our work aims to promote more rigorous and interpretable statistical practice in neurological trials.

Poster Highlights

Monday, 2025-08-25 11:30 - 13:00, Biozentrum U1.141

Chairs: Jenny Devenport and Jianmei Wang

Get a rapid overview of 30 innovative posters in this high-energy session, where each presenter delivers a 2-minute teaser of their work. From novel statistical methods to impactful clinical applications, this showcase offers a fast-track to the most exciting developments in clinical biostatistics. Perfect for planning your poster tour and sparking new collaborations.

1. **Monday, BioZ 38,**
Fukuyama, Yuki:
Usefulness of the blinded sample size re-estimation for dose-response trials with MCP-Mod
2. **Monday, BioZ 28,**
Dardenne, Nadia:
Investigating (bio)statistical literacy among health researchers in a Belgian university context: A framework and study protocol
3. **Monday, BioZ 27,**
Sasaki, Kotaro:
A novel approach for assessing inconsistency in network meta-analysis: Application to comparative effectiveness analysis of antihypertensive treatments
4. **Tuesday, ETH 25,**
Chernova, Oksana:
Building cancer risk prediction models by synthesizing national registry and prevention trial data
5. **Tuesday, BioZ 9,**
WANG, ZIYAN:
Estimands in platform trials with time - treatment interactions
6. **Tuesday, ETH 24,**
Jeon, Somin:
Refining the Association between BMI, Waist Circumference, and Breast Cancer Risk in Postmenopausal Women using G-formula Method
7. **Monday, ETH 9,**
Trutschel, Diana:

Statistical Approach to Assess the Impact of Hospital Settings on Optimal Staffing Levels

8. **Tuesday, BioZ 12,**

Musik, Szymon:

Leveraging Synthetic Data for Enhanced Clinical Research Outcomes

9. **Tuesday, BioZ 2,**

Lee Alcober, Maria:

Development and validation of prognostic models in phase I oncology clinical trials

10. **Monday, BioZ 25,**

Shi, Qingyang:

Learning heterogeneous treatment effect from multiple randomized trials to inform healthcare decision-making: implications and estimation methods

11. **Tuesday, BioZ 40,**

Ambrosetti, Francesco:

Feasibility of propensity score weighted analysis in rare disease trials: a simulation study

12. **Monday, ETH 20,**

Esteban, Luis Mariano:

Balancing Accuracy, Clinical Utility, and Explainability: A Machine Learning Approach to Prostate Cancer Prediction

13. **Monday, ETH 34,**

Lee, Jinwoo:

Tree-based methods for length-biased survival data

14. **Monday, BioZ 5,**

Sabanés Bové, Daniel:

Designing Clinical Trials in R with ract and crmPack

15. **Tuesday, ETH 26,**

White, Bethan L.:

Modelling Individual-level Uncertainty from Missing Data in Personalised Breast Cancer Risk Prediction

16. **Monday, ETH 17,**

Huybrechts, Amber:

Cell composition analysis with unmeasured confounding

17. **Tuesday, ETH 32,**
Pamminger, Moritz:
Long-term risk prediction from short-term data, a microsimulation approach
18. **Tuesday, BioZ 23,**
Ahmad, Mahru:
Reducing Uncertainty in Fertility Meta-Analysis: A Multivariate Approach to Clinical Pregnancy and Live Birth Outcomes
19. **Tuesday, BioZ 15,**
Potapov, Ilya:
Strategies to scale up model selection for analysis of proteomic datasets using multiple linear mixed-effect models
20. **Monday, ETH 33,**
Štěpánek, Lubomír:
Non-parametric methods for comparing survival functions with censored data: Exhaustive simulation of all possible beyond-observed censoring scenarios and computational analysis
21. **Monday, BioZ 43,**
Jelizarow, Monika:
A pre-study look into post-study knowledge: communicating the use(fulness) of pre-posteriors in early development design discussions
22. **Tuesday, ETH 21,**
Bobek, Olivia J:
Leveraging tumor imaging compositional data structure in model feature space for predicting recurrence in colorectal carcinoma
23. **Tuesday, ETH 23,**
Dennis, Divya:
Joint Modelling of Random Heterogeneity in Longitudinal and Multiple time-to-Events in Colon Cancer
24. **Tuesday, ETH 33,**
Margaté, Tristan:
Deep learning algorithm for dynamic survival prediction with competitive risks
25. **Tuesday, BioZ 29,**
Liang, Xiaoran:
Assessing the effect of drug adherence on longitudinal clinical outcomes: A comparison

of Instrumental Variable and Inverse Probability Weighting methods

26. **Monday, BioZ 45,**
Schneider, Juliana:
Dealing with missing values in adaptive N-of-1 trials
27. **Tuesday, ETH 5,**
Cornett, Chantelle:
A Systematic Review of Methodological Research on Multi-State Prediction Models
28. **Monday, BioZ 46,**
Salsbury, James:
Adaptive clinical trial design with delayed treatment effects using elicited prior distributions
29. **Tuesday, ETH 12,**
Zhai, Yue:
Marginal structural Cox model with weighted cumulative exposure modelling for the estimation of counterfactual Population Attributable Fractions
30. **Monday, BioZ 7,**
Castagné, Claire:
Design of a research project to evaluate the statistical utility after transformation of a CDISC database into OMOP format

Dynamic Borrowing and Basket Trials

Monday, 2025-08-25 11:30 - 13:00, Biozentrum U1.101

Chair: Isaac Gravestock

1: Utility-Based Optimization of Basket Trials

Lukas D Sauer¹, Alexander Ritz², Michaela Maria Freitag³, Meinhard Kieser¹

¹Institute of Medical Biometry, Heidelberg University, Germany

²Institute of Mathematics, Clausthal University of Technology, Germany

³Institute of Biometry and Clinical Epidemiology, Charité – Berlin University of Medicine, Germany

Introduction The dawn of personalized medicine comes with new challenges in the design of clinical trials. Especially in the development of gene-specific cancer therapy, drugs may be tissue-agnostic, which means that the same drug can be applied in diseases of different organs. So-called *basket trial designs* are a useful tool for investigating such drugs: In a single trial, one therapy is investigated in several strata defined by different diseases or disease subtypes. This eases the organizational burden compared to the conduct of several trials, and furthermore it offers a statistical benefit: *information borrowing* can be used to share response information between the strata. This may leverage the oftentimes small sample sizes in the strata to increase power while keeping type-I error inflation and bias reasonably low. In biometrical research, a variety of frequentist and Bayesian techniques were suggested to implement borrowing in basket trials. These designs often come with tuning parameters to adjust the amount of borrowing. However, the optimal choice of these parameters is subject to current research.

Methods We suggest a utility-based framework for optimizing borrowing in basket trial designs. As an example, we consider a Bayesian basket trial design by Fujikawa et al. (Biom J, 2020;62(2):330–8) which is based on the beta-binomial model. We demonstrate the use of utility functions for defining a compromise between successfully detecting strata with high response rates (“*active strata*”) while rejecting strata with low response rates (“*inactive strata*”). Applying numerical optimization algorithms, these utility functions can be used to find optimal tuning parameters for the borrowing.

Results and conclusion We conducted an extensive comparison study showing that utility-based optimization is a feasible approach for optimizing information borrowing in basket trial designs. Our approach successfully achieves a compromise between increasing per-stratum

power while keeping type-I error inflation moderate. In our study, we also compared the use of grid search to different optimization algorithms and will discuss their performance. Our framework is not unique to Fujikawa's design and may hence be used for planning basket trial designs in general. Finally, we will also discuss the extension of utility functions to tune sample size and interim analyses, thus allowing simultaneous planning of all aspects of basket trial design in a single framework.

2: A Frequentist Approach to Dynamic Borrowing

Ray Lin¹, Ruilin Li², Jiangeng Huang¹, Lu Tian², Jiawen Zhu¹

¹Roche, United States of America

²Stanford University

Background

There has been growing interest in leveraging external control data to augment a randomized control group data in clinical trials and enable more informative decision making. In recent years, the quality and availability of real-world data have improved steadily as external controls. However, information borrowing by directly pooling such external controls with randomized controls may lead to biased estimates of the treatment effect. Dynamic borrowing methods under the Bayesian framework have been proposed to better control the false positive error. However, the numerical computation and, especially, parameter tuning, of those Bayesian dynamic borrowing methods remain a challenge in practice.

Method

We present a frequentist interpretation of a Bayesian commensurate prior borrowing approach and describe intrinsic challenges associated with this method from the perspective of optimization. Motivated by this observation, we propose a new dynamic borrowing approach using adaptive lasso. The treatment effect estimate derived from this method follows a known asymptotic distribution, which can be used to construct confidence intervals and conduct hypothesis tests. The proposed approach only needs to solve two convex optimizations problems and thus is easy to implement and computationally efficient. Extensive Monte Carlo simulations were conducted under different settings to evaluate the type I error, power, bias, standard deviation, mean square error (MSE) of treatment effect estimates and effective sample size (ESS).

Results

We observed highly competitive performance of adaptive lasso compared to Bayesian approaches, in terms of statistical power, bias, MSE, and ESS. An illustration example using actual trial data also suggested the adaptive lasso produces similar estimates as the Bayesian approaches, yet with much less computation time and resources. Methods for parameter tuning are also thoroughly discussed based on results from the stimulation studies.

Conclusion

We have developed a novel frequentist dynamic borrowing method based on the adaptive lasso methodology. This approach boosts the accuracy of treatment effect estimation by borrowing information from external controls. It is easy to implement, runs substantially faster than Bayesian approaches, and allows frequentist asymptotic inference. While it is theoretically impossible to have an estimator that always outperforms a Bayesian estimator (and vice versa), we have demonstrated through extensive simulations that with appropriate choices of tuning parameters, the operating characteristics of our approach is highly competitive compared to conventional Bayesian approaches.

3: A Power Prior Based Basket Trial Design for Unequal Sample Sizes

Sabrina Schmitt^{1,2}, Lukas Baumann¹

¹University of Heidelberg, Germany

²University of Würzburg, Germany

Not presented.

Background Basket trials examine the efficacy of a single intervention simultaneously in several patient subgroups, called baskets. They are currently mostly applied in oncology, where the assignment to the baskets is based on matching medical characteristics such as a common mutation. This can result in small sample sizes within baskets that are also likely to differ. 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 individual 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 trial designs with respect to the expected number of correct decisions, type 1 error rates and power. We consider a range of different scenarios with different true response probabilities and sample sizes across baskets.

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.

Conclusion 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.

4: Borrowing Information in Basket Trials with Different Clinical Outcomes via a Common Intermediate Outcome

Svetlana Cherlin, James M S Wason

Newcastle University, United Kingdom

Basket trials are a new class of trial designs that evaluate a common treatment across multiple related conditions. These trials allow for more efficient analysis by leveraging information from different subtrials; however, methods typically assume a common endpoint. In immune-mediated inflammatory disease trials, different clinical trial endpoints are often used, making direct information sharing between subtrials in a basket setting less justifiable. However, in these diseases, the response to treatment is often mediated through a common inflammatory biomarker. For example, in trials investigating the safety and efficacy of Ustekinumab for Ulcerative Colitis, clinical remission was the primary endpoint, whereas in trials of the same drug for Rheumatoid Arthritis, the primary endpoint was the disease activity score. These

trials also collected data on C-reactive protein, which may mediate the treatment effect on clinical outcomes. It is plausible that borrowing information on the clinical outcome treatment effect via the mediator would enhance the efficiency of the analysis. We develop methodology that borrows information on the mediator in basket trials with distinct responder outcomes.

We propose a Bayesian hierarchical model that assumes the treatment affects an outcome both directly and indirectly through an intervening mediator variable. The model allows for borrowing of information between subtrials, with the extent of borrowing determined by the prior distributions of the parameters for the mediation effect. Since the outcomes are assumed to be different, there is no sharing of information on the direct treatment effect on the outcome; the only sharing occurs through the mediator. We investigate the operating characteristics of the model using a simulation study and apply it to real data from Ustekinumab trials.

We compared the new approach with logistic regression on a binary outcome that shares information on the response outcome, across multiple simulation scenarios. In scenarios with a mediation effect, we observed an increase in power and a reduction in the width of credible intervals for log odds ratios provided by the new model. In many scenarios, the new model achieved a decrease in bias and mean squared error, as well as an increase in precision for the estimates. The type I error rate was well controlled.

Numerical results suggest that sharing information on the mediation effect can improve the precision of estimates and increase power compared to a standard approach. Further work will explore the most effective mechanism for sharing information between the subtrials through suitable prior distributions of the parameters.

5: Robust External Information Borrowing in Hybrid-Control Clinical Trial Designs

Silvia Calderazzo¹, Manuel Wiesenfarth², Vivienn Weru¹, Annette Kopp-Schneider¹

¹German Cancer Research Center, Germany

²Cogitars, Germany

Introduction External information borrowing can help improving clinical trial efficiency and is therefore often considered in situations where the sample size that can realistically be recruited is limited, as, e.g., pediatric or rare disease trials. When dealing with a two-arm trial, external information is often available for the control arm, leading to so-called 'hybrid-control' designs. The Bayesian approach borrows such external information by adopting an informative prior distribution for the control arm mean. A potential issue of this procedure

is that external and current information may conflict, but such inconsistency may not be predictable a priori. Robust prior choices are typically proposed to limit extreme worsening of operating characteristics in these situations. However, trade-offs are still present and in general no power gains are possible if strict control of type I error (TIE) rate is desired. In this context, principled justifications for TIE rate inflation can be of interest.

Methods Building on Calderazzo et al. (2024), we propose an interpretable approach for external information borrowing in this context. The method is built by eliciting explicit caps on TIE rate inflation and power loss based on regions of pre-defined conflict. The method does not rely on a prior specification, and is developed for both normal and binomial outcomes by exploiting relationships between frequentist and Bayesian test decisions thresholds.

Results The method provides a direct and easily interpretable link between conflict assumptions and test error rates behavior. We additionally show connections of the approach to alternative robust borrowing methods, which can aid their interpretability. The long run behavior (including Bayesian average TIE rate and power) is evaluated via simulations, showing comparable performance to alternative dynamic borrowing methods, such as the robust mixture prior.

Conclusions Robust external information borrowing typically depends on the choice or estimation of specific borrowing parameters. We provide a rationale for robust incorporation of external information which is directly linked to TIE rate inflation, thus helping interpretability of such borrowing parameters.

Calderazzo, S., Wiesenfarth, M., & Kopp-Schneider, A. (2024). Robust incorporation of historical information with known type I error rate inflation. *Biometrical Journal*, 66(1), 2200322.

Causal Inference - Dealing with Bias

Monday, 2025-08-25 11:30 - 13:00, ETH E27

Chair: Julie Josse

1: Proximal Indirect Comparison

Zehao Su¹, Helene Rytgaard¹, Henrik Ravn², Frank Eriksson¹

¹University of Copenhagen

²Novo Nordisk

Background We consider the problem of indirect comparison, where a treatment arm of interest is absent by design in one randomized controlled trial (RCT) but available in the other. That is, we are interested in the contrast of treatments that are not compared in head-to-head RCTs. e.g., the comparison of a new treatment and an existing treatment, when both treatments have only been studied in placebo-controlled RCTs.

Identifiability of the target RCT population average treatment effect often relies on conditional transportability assumptions. However, it is a common concern whether all relevant effect modifiers are measured and controlled for. If the treatments of interest come from RCTs which are conducted with a considerable time gap apart, there may be a drift in unmeasured social determinants of health or changes in the standard of care or that could affect the treatment effects. When there are unobserved shifted effect modifiers, transportability cannot be established by controlling for observed baseline variables.

Methods and results Recently a family of methods called proximal causal inference has shown how appropriately selected proxies, or negative controls, may rectify confounding bias. We borrow these ideas and propose a novel method that uses proxies to tackle bias arising from shifted unobserved effect modifiers. We give a new proximal identification result based on two proxies, an adjustment proxy in both RCTs and an additional reweighting proxy in the source RCT. We propose an estimator that is doubly-robust against misspecifications of the so-called bridge functions and asymptotically normal under mild consistency of estimators for the bridge functions.

We use two placebo-controlled weight management trials conducted 5-6 years apart as a context to illustrate selection of proxies and apply our method to compare the weight loss effect of the active treatments from these trials.

Conclusion Proximal indirect comparison may allow for treatment effect estimation via transportability in the presence of unobserved effect modifiers. The proximal indirect comparison estimator can be bias-free even when transportability of the conditional average treatment effect fails to hold conditioning on the observed data.

2: Selection Bias of Cause-Specific Hazard Ratios: The Impact of Competing Events

Mari Brathovde^{1,2}, Morten Valberg^{1,2}, Hein Putter³, Richard A.J. Post⁴

¹Oslo University Hospital, Norway

²University of Oslo, Norway

³Leiden University Medical Center, The Netherlands

⁴Erasmus Medical Center, The Netherlands

Background Competing risks generalize standard survival analysis of a single, often composite outcome when interest lies in the different causes of the event. In the presence of heterogeneity, the complex causal interpretation of the hazard ratio for all-cause mortality is well-known and has been formalized. Yet the current recommendation in epidemiology is to use cause-specific hazard ratios when aiming to understand etiology in a competing risk setting.

Methods In this work, we formalize how observed cause-specific hazard ratios evolve and deviate from the (conditional) causal effect of interest in the presence of heterogeneity of the hazard rate of unexposed individuals (frailty) and heterogeneity in effect (individual modification). Importantly, we do so without imposing assumptions on the baseline hazard rate or the distributions of the latent effect modifiers and frailty factors. We show that the presence of a competing event can amplify the selection bias of the all-cause mortality setting, as it introduces selection on the frailties and effect modifiers associated with both the event of interest and the competing event.

Results We provide illustrative examples using frailties from the family of power variance function (PVF) distributions, along with categorical effect modifiers (harmful, beneficial, or neutral). The PVF family of frailties yields convenient analytical expressions for the observed cause-specific hazard ratios, enabling a clear separation of the selection bias introduced by different events. This approach allows for a straightforward evaluation of the impact of treatment effects and event prevalence on the bias using simple monotonicity principles. The

numerical examples include settings with crossover of the cause-specific hazard rates between exposed and non-exposed individuals, which would not occur without competing events.

Conclusion We show that the size and sign of the bias in the cause-specific hazard ratios depend on the prevalence of, and the treatment's effect on, the competing event. Consequently, the cause-specific hazard ratio suffers from both selection bias inherent to all-cause mortality hazard ratios and dependence on the competing event, complicating its causal interpretation and limiting its suitability for addressing etiological questions without relying on untestable assumptions. This work highlights the importance of employing more appropriate estimands in a competing risk setting, such as marginal cumulative incidences.

3: Qbaconfound: A Flexible Monte Carlo Probabilistic Bias Analysis to Unmeasured Confounding

Emily Kawabata, Chin Yang Shapland, Tom Palmer, David Carslake, Kate Tilling, Rachael Hughes

University of Bristol

Background Unmeasured confounding is a persistent concern in observational studies. We can quantitatively assess the impact of unmeasured confounding using a probabilistic bias analysis (PBA). A PBA specifies the relationship between the unmeasured confounder(s), U , and study data via its bias parameters. External information about U is incorporated via prior distribution(s) placed on these bias parameters. A Bayesian PBA combines the prior distribution(s) with the data's likelihood function whilst a Monte Carlo PBA samples the bias parameters directly from its prior distributions. Software implementations of PBAs to unmeasured confounding are scarce and mainly limited to unadjusted analyses of a binary exposure and outcome. One exception is R package *unmconf* (Hebdon et al 2024, BMC Med. Res. Methodol., <https://doi.org/10.1186/s12874-024-02322-2>) which implements a Bayesian PBA, applicable when the analysis is a generalised linear model (GLM). However, for a study with P measured confounders and a single U , *unmconf* requires information on $3+P$ to $6+2P$ bias parameters, which is burdensome when validation data are unavailable.

Aim We propose a flexible Monte Carlo PBA where the number of bias parameters is independent of the number of measured confounders. It is applicable to a GLM or survival proportional hazards model, with binary, continuous, or categorical exposure and measured confounders, allows for nonlinear and interaction terms (of exposure and measured confounders), and one or more binary or continuous unmeasured confounders.

Methods Via simulations, we evaluate our PBA for different analyses (e.g., varying the regression model, exposure type, and with or without interactions), and different levels of dependency between the measured and unmeasured confounders. Also, we compare our Monte Carlo approach to a fully Bayesian implementation when fitting a linear or logistic regression. We repeat the simulation study for prior distributions with different levels of informativeness.

Results Ignoring U resulted in substantially biased estimates with substantial confidence interval undercoverage (e.g., 56%). Our Monte Carlo PBA (with informative priors) results in point estimates with minimal or no bias and interval estimates with close to nominal coverage. For binary U, levels of bias are marginally higher when U is strongly correlated with the measured confounders. The performances of the Monte Carlo and Bayesian implementations are comparable except the Monte Carlo version is slightly quicker.

Conclusion Our Monte Carlo PBA is applicable to a variety of regression-based analyses, with minimal burden to the user. It will be implemented as a Stata command and R package, *qbaconfound*.

4: Selection Bias Due to Omitting Interactions from Inverse Probability Weighting

Liping Wen, Kate Tilling, Rosie Cornish, Rachael Hughes, Apostolos Gkatzionis

University of Bristol, United Kingdom

Background The estimated causal effect of an exposure on an outcome might be biased if the analysis sample is subject to selection, e.g. due to non-random participation or dropout. Inverse probability weighting (IPW) is often used to adjust for selection bias, typically using a simple logit weighting model without interactions. However, the size of selection bias depends on the interaction between exposure and outcome in their effect on selection. This implies that it may be important to include interaction terms in the IPW model.

Methods We compare IPW methods using a simulation study, where the estimand of interest is the causal effect of exposure on outcome. The selection mechanisms are simulated using a logit/log-additive/probit model with and without interaction terms, with either exposure, causes of exposure, or causes of outcome influencing selection. We analysed each simulated dataset by: full data analysis, complete case analysis (CCA), and four IPW methods which differ only in the weighting model used. These weighting models are either a logit or log-additive model, and either include or exclude the interaction between variables causing the

selection. In a real-data application, we use data from the Understanding Society study to investigate the effect of unemployment on sleep duration, using IPW to adjust for dropout. Exposure, mediator and 6 confounders, as well as all possible two-way interactions, are included in the weighting model. Then the least absolute selection and shrinkage operator (LASSO) is used for variable selection to avoid overfitting.

Results The simulation study shows that IPW including an interaction term gives less biased estimates than IPW without this term in all scenarios studied. Importantly, IPW using a logit model with no interaction terms often gives estimates very close to CCA. IPW including 89 interaction terms suggests that unemployment reduces sleep duration by around 23 (9, 38) minutes, compared to 27 (14, 40) minutes for IPW without interactions, and 31 (19, 43) minutes for the CCA.

Conclusion We strongly recommend that researchers include interaction terms in weighting models to adjust for selection bias. Also, an agreement between CCA and IPW without interactions arises mathematically and so does not indicate that results are robust to selection bias.

5: External Reproduction of a Proxy-Based Causal Model Estimating Average Survival Effects of Sequential vs Concurrent Chemo-Radiotherapy in Stage III NSCLC

Charlie Cunniffe¹, Wouter van Amsterdam², Matthew Sperrin¹, Rajesh Ranganath³, Fiona Blackhall^{1,4}, Gareth Price¹

¹The University of Manchester, United Kingdom

²University Medical Center Utrecht, Netherlands

³New York University, USA

⁴The Christie NHS Foundation Trust, United Kingdom

Background/Introduction Randomised controlled trials (RCTs) offer the most reliable evidence for treatment decisions. However, in cancer care, frail, elderly, and disadvantaged patients are often underrepresented, causing uncertainty about optimal treatment strategies, such as whether sequential or concurrent chemo-radiotherapy yields better outcomes. Observational causal inference may supplement RCT evidence; however, confounding factors that are not directly observed remain a challenge. Using available proxy measurements to infer these variables offers a potential solution. In the cancer setting, the patient's overall fitness is an important unobserved confounder for treatment and outcome, for which proxies like "performance score" are widely available within patient records. This study reproduces

a recently introduced proxy-based causal inference method to assess transportability over populations with different data structures and treatment guidelines.

Methods This study employs proxy-based individual treatment effect modelling in cancer (PROTECT) to estimate the individual treatment effect of concurrent vs sequential chemo-radiotherapy on overall survival in 1117 routinely treated patients with stage III Non-Small Cell Lung Cancer (NSCLC) seen between 2013 and 2023. A local model was developed by adapting the PROTECT directed acyclic graph to include our selected proxies of patient fitness – performance, comorbidity, and frailty scores. While the causal effect is generally not identifiable in the presence of unobserved confounding, PROTECT assumes that the proxies are conditionally independent given the unobserved confounder and incorporates domain knowledge via functional constraints on a Bayesian latent factor model. Individual treatment effect estimates for each patient are averaged to get the population's average treatment effect (ATE) estimate, reported as a hazard ratio. We compared the ATE to standard multivariable Cox regression, the ATE from the PROTECT development study and the results of a meta-analysis of RCTs.

Results The ATE on survival in our population is 0.92 [0.73, 1.15] in favour of concurrent chemo-radiotherapy. The standard multivariable analysis gives 0.60 [0.48, 0.75]. The PROTECT development publication reports 1.01 [0.68, 1.53]. The RCT meta-analysis reports a stronger effect size (0.84 [0.74, 0.95]) than the causal analysis but has a younger (19% > 70 vs 43%) and fitter (50% ECOG 0 vs 27%) population than our study.

Conclusion Our results show causal analysis yields more credible ATEs than standard methods. RCT inclusion criteria explain differences from meta-analysis ATEs. Repeated analysis in two separate populations yielded comparable results, implying a robust estimation of the ATE. PROTECT is a promising new method that can be applied more broadly in cancer and other settings.

Survival Analysis 1

Monday, 2025-08-25 11:30 - 13:00, ETH E23

Chair: Thomas Scheike

1: Standardized Survival Probabilities and Contrasts Between Hierarchical Units in Multilevel Survival Models

Alessandro Gasparini¹, Michael J. Crowther¹, Justin M. Schaffer²

¹Red Door Analytics AB, Sweden

²Department of Cardiothoracic Surgery, Baylor Scott & White The Heart Hospital, Plano, Texas, USA

Observational data with time-to-event (survival) outcomes in medical research often exhibit a hierarchical (or clustered) structure. For instance, siblings tend to be more similar than randomly selected individuals from the general population, and study subjects may be nested within geographical areas or institutions such as schools and hospitals. Subjects within the same hierarchical unit are likely correlated, and ignoring this multilevel structure can lead to biased and inefficient results.

Multilevel hierarchical mixed-effects survival models are commonly used in such settings: these models account for correlation among study subjects within the same cluster and any potential unobserved heterogeneity. However, such analyses typically focus on fixed effects while marginalizing over the random effects, leading to post-estimation predictions that have either a subject-specific or population-level interpretation. Nevertheless, comparing hierarchical units is crucial in certain situations - such as when benchmarking the performance of medical institutions and providers while accounting for differences in case-mix covariates, when quantifying heterogeneity between clusters, or when comparing trials included in individual patient data (IPD) meta-analysis.

In this work, we propose combining regression standardization with cluster-specific posterior predictions of the random effects to quantify the performance of each hierarchical unit. By fixing the predicted random effects and standardizing over the remaining (observed) covariates, we obtain model-based predictions that retain their usual interpretation as survival probabilities - either at a specific time point or over the entire follow-up period. With multiple hierarchical levels, we can also isolate the effect of a specific level while marginalizing over the others. Contrasts of standardized survival probabilities can then be computed, which retain the natural interpretation as risk ratios or differences.

These standardized predictions quantify how the entire study population would have fared under the performance of each cluster, enabling fair comparisons between hierarchical units. Compared to commonly used approaches for quantifying contextual effects, such as the median hazard ratio, our proposed method yields more interpretable quantities that, under certain assumptions, can also have a causal interpretation.

We illustrate this methodology using data on bladder cancer patients with three levels of nesting: patients within surgeons, and surgeons within centers. Finally, we developed user-friendly software in Stata and R to facilitate the application of this method in practice.

2: Multimodal Mixture Regression on Censored Data with a Cure Fraction.

Mathilde Foulon, Anouar El Ghouch, Catherine Legrand

UCLouvain, Belgium

Introduction

There is an abundant literature in statistics, biostatistics and econometrics on the modelling, estimation and inference of regression models for survival data subject to censoring. However, only a few of them consider a potential multimodality of the time-to-event. To the best of our knowledge, there is no model that considers both multimodality and the possible presence of a cure fraction, i.e. the presence of a fraction of subjects who do not experience the event of interest. Our aim is to develop a modelling approach that takes both these aspects into account. This is particularly useful in contexts such as modelling cancer recurrence, where recurrences may occur in several waves, but with a proportion of patients never relapsing.

Methods

In this work, we have built a model that considers both multimodality of the time-to-event and a cure fraction. To achieve this aim, we developed an accelerated failure time model in which the error term is assumed to follow a mixture of Sinh-Cauchy distributions. This approach offers greater robustness by combining the flexibility of mixture models with that of the Sinh-Cauchy distribution. We studied the properties of this distribution and implemented an estimation method using the EM algorithm. A simulation study was carried out to illustrate the performance of the proposed approach.

Results

Simulations have demonstrated the relevance and effectiveness of our approach for modelling multimodal time-to-event data with a proportion of cure. The methodology implemented to estimate the various parameters of the model provides reliable results in terms of bias, variance and MSE. Further investigations are ongoing on the selection of the number of components in the mixture, but preliminary results indicate that large flexibility is already achieved with a limited number of mixture components.

Conclusion

The results obtained show that the proposed model is an interesting alternative to traditional cure models in the presence of multimodality and cure while also providing good results for unimodal data. It therefore constitutes a more flexible and robust approach in the context of multimodal survival data with a proportion of cure. In the following, we intend to apply our methodology to real data.

3: Modelling & Assessing the Effect of Frailty and Longitudinal Measures in Time-to-Event Outcome: An Application to Colorectal Cancers

Anand Hari, Jagathnath Krishna Km

Regional Cancer Centre, India

Introduction Joint modelling is a powerful statistical framework enabling a simultaneous modelling of longitudinal covariate and time-to-event outcome. Typically, a joint model consist of two subparts, a longitudinal sub-model which is modelled using a linear mixed model and a survival sub-model by Cox-proportional hazard model or parametric survival models. Apart from the interdependence between longitudinal and survival outcomes there may exist some unobserved random heterogeneity which can be modelled using a joint frailty model.

Methods Here we considered joint models to determine the unobserved heterogeneity among individuals by integrating frailty terms in the survival sub-model. Incorporating frailty in the survival sub-model can allow an additional source of variation at the survival endpoint that cannot be explained by the longitudinal data. So here we developed joint frailty model and compared the results with Joint model, Cox model and Frailty model. The analyses were performed using R Software and illustrated using colorectal cancer data.

Results Joint frailty model demonstrated improved model fit compared to the conventional Cox and Frailty model. When compared with the joint model, the joint frailty model accounts

for frailty among individuals or clusters. The inclusion of frailty in the survival sub-model captured additional unexplained heterogeneity among individuals, leading to more precise hazard estimates. The association between the longitudinal biomarker and survival outcome and the frailty variance was found to be statistically significant, reinforcing the benefit of joint frailty modelling. Additionally, the significance of variance of the frailty term suggested the presence of substantial unobserved heterogeneity among individuals.

Conclusion The joint frailty model provided a more comprehensive framework for analysing longitudinal and survival data by accounting for both the interdependence between these outcomes and unobserved individual variability. The findings highlight the importance of incorporating frailty terms in joint modelling, particularly in cancer prognostic studies where individual heterogeneity plays a crucial role.

Keywords Joint model, Cox-Proportional hazard, Frailty, Joint-frailty model, Colorectal Cancer

4: Adjusted Kaplan-Meier Curves for Partly Unobserved Group Membership in Paediatric Stem Cell Transplantation Studies

Martina Mittlböck¹, Harald Heinzl¹, Ulrike Pötschger²

¹CeDAS, Medical University of Vienna, Austria

²Children's Cancer Research Institute, Austria

Background Randomisation is hardly achievable in controlled clinical trials of childhood leukaemia comparing standard continued chemotherapy with allogeneic stem cell transplantation (SCT). Usually standard chemotherapy will be stopped and SCT performed if a donor search identifies a suitable stem cell donor in existing registries of potential donors.

A fair comparison can be based on the two groups formed by the availability or non-availability of suitable stem cell donors. Donor availability is a temporarily unknown external baseline variable whose actual values will become known either when a suitable stem cell donor is identified from the registry or after the end of an unsuccessful donor search. However, donor search can be prematurely ceased, e.g. due to patients' death or deterioration of patients' health. Unfortunately, then donor availability and thus group membership will remain unknown. There is currently no graphical approach available to correctly illustrate survival probabilities over time for the two groups.

Methods For each patient with prematurely ceased donor search, it is possible to calculate

the probabilities that a suitable donor might or might not have been identified after the ceasing of the donor search, respectively. These probabilities are utilized to develop adjusted Kaplan-Meier curves for visual group comparison over time. These curves have a valid survival probability interpretation unlike the commonly applied Simon and Makuch curves where patients are allowed to change groups over time.

Estimated survival probabilities derived from adjusted Kaplan-Meier curves can also be used to assess group differences at selected long-term time points, which is especially interesting in case of non-proportional hazards. A corresponding statistical test is proposed and compared to other test strategies.

Results Data from an international study of children with newly diagnosed Philadelphia chromosome-positive acute lymphoblastic leukaemia are used to exemplify the satisfactory performance of the new approach. Other methods are only able to estimate survival probabilities at selected time points. Results of the different methods are compared and discussed.

Conclusion The newly proposed method allows for the first time to show Kaplan-Meier curves, when group membership at baseline is unknown and becomes only partly known over time.

5: Extended Multi-Stage Drop-the-Losers Design for Multi-Arm Clinical Trials using Binary and Survival Endpoints

Manuel Pfister^{1,2}, Pierre Colin²

¹University of Zurich, Division of Biostatistics and Reproducibility, Zurich, Switzerland

²Bristol Myers Squibb, Global Biometrics & Data Sciences, Boudry, Switzerland

In oncology, clinical trials are a cornerstone of evaluating new treatments. However, traditional designs often face significant challenges in efficiency, particularly when assessing multiple treatment options. The emergence of Multi-Arm Multi-Stage (MAMS) designs, such as the "Drop-the-losers" approach, offers innovative solutions by combining multiple hypotheses within a single trial framework.

This work evaluates the "Drop-the-losers" approach (as published by J. Wason et al. in 2017), focusing on its application in Phase II/III oncology trials. We extend existing methodologies by incorporating binary and survival endpoints, addressing limitations of original design which focused primarily on continuous outcomes. Using simulation studies, the statistical performance of the "Drop-the-losers" design was assessed across various scenarios. Key metrics

included Type I error control, statistical power, and biases under varying endpoint distributions. Simulations involved a survival endpoint such as progression-free survival (PFS). Results demonstrated that the "Drop-the-losers" design effectively balances statistical rigor and efficiency. The design strongly controls Type I error while ensuring high power in detecting drug effects. Scenarios with harmful or ineffective treatments highlighted the ethical advantage of eliminating suboptimal options early, minimizing patient exposure to ineffective therapies.

One limitation identified is that long-term survival endpoints may not be mature enough to support early treatment selection analyses. Updating the endpoint of interest over time could align with clinical practices, starting with ORR at the first analysis, transitioning to Benefit-Risk scores or PFS at the second analysis, and concluding with OS at the final analysis. However, such an approach requires careful handling of correlations between drug effects across these endpoints (i.e., surrogacy).

This work provides a comprehensive framework for implementing "Drop-the-losers" design in clinical trials, demonstrating its potential to accelerate treatment evaluation in oncology while upholding ethical and scientific standards. This methodology contributes to the ongoing evolution of designs, underscoring their value in modern clinical research.

Wason, James et al. "A multi-stage drop-the-losers design for multi-arm clinical trials." *Statistical methods in medical research* vol. 26,1 (2017): 508-524. doi:10.1177/0962280214550759

Prediction / Prognostic Modelling 1

Monday, 2025-08-25 11:30 - 13:00, ETH E21

Chair: Aris Perperoglou

1: Developing a Clinical Prediction Model with a Continuous Outcome: Sample Size Calculations to Target Precise Predictions

Rebecca Whittle^{1,2}, Richard D. Riley^{1,2}, Lucinda Archer^{1,2}, Gary S. Collins³, Paula Dhiman³, Amardeep Legha^{1,2}, Kym Snell^{1,2}, Joie Ensor^{1,2}

¹Department of Applied Health Sciences, School of Health Sciences, College of Medicine and Health, University of Birmingham, United Kingdom

²National Institute for Health and Care Research (NIHR) Birmingham Biomedical Research Centre, United Kingdom

³Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, United Kingdom

Background When developing a clinical prediction model, the precision of predictions is heavily influenced by the sample size used for development. Without adequate sample sizes, models may yield predictions that are too imprecise to usefully guide clinical decisions. Previous sample size research for developing models with a continuous outcome is based on minimising overfitting and targeting precise estimation of the residual standard deviation and model intercept. However, even when meeting these criteria, the uncertainty (instability) in predictions is often considerable. We propose a new approach for calculating the sample size required to target precise individual-level predictions when developing a prediction model for a continuous outcome.

Methods We outline a four-step approach which can be used either before data collection (based on published aggregate data), or when an existing dataset is available (e.g., from a pilot study or existing study/database). We derive closed-form solutions that decompose the anticipated variance of individual outcome estimates into Fisher's unit information matrix, predictor values and total sample size.

Results The approach allows researchers to examine anticipated interval widths of individual predictions based on one particular sample size (i.e., of a known existing dataset), or to identify the sample size needed for a new study aiming to target a certain level of precision (e.g., a new cohort study). Additionally, this can be examined in particular subgroups of patients to help improve fairness of the model. We use a real example predicting Forced Expiratory

Volume (FEV) in children to showcase how the approach allows researchers to calculate and examine expected individual-level uncertainty interval widths for particular sample sizes. We also showcase our new software module *pmstabilityss*.

Conclusions We derived a new approach to determine the minimum required sample size to develop a clinical prediction model with a continuous outcome that gives precise individual outcomes. The approach enables researchers to assess the impact of sample size on the individual-level uncertainty; to calculate the required sample size based on a specified acceptable level of uncertainty; and to examine differences in precision across subgroups to inform fairness checks.

2: Sequential Sample Size Calculations for Developing Clinical Prediction Models: Learning Curves Suggest Larger Datasets are Needed for Individual-Level Stability

Amardeep Legha^{1,2}, Joie Ensor^{1,2}, Ben Van Calster^{3,4}, Evangelia Christodoulou⁵, Lucinda Archer^{1,2}, Rebecca Whittle^{1,2}, Kym I.E. Snell^{1,2}, Paula Dhiman⁶, Gary S. Collins⁶, Richard D. Riley^{1,2}

¹Department of Applied Health Sciences, School of Health Sciences, College of Medicine and Health, University of Birmingham, Birmingham, United Kingdom

²National Institute for Health and Care Research (NIHR) Birmingham Biomedical Research Centre, United Kingdom

³Leuven Unit for Health Technology Assessment Research (LUHTAR), KU Leuven, Leuven, Belgium

⁴Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the Netherlands

⁵German Cancer Research Center (DKFZ) Heidelberg, Division of Intelligent Medical Systems, Heidelberg, Germany

⁶Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, United Kingdom

Background Clinical prediction models (CPMs) estimate an individual's risk of a particular outcome to inform clinical decision-making. Small sample sizes may lead to unreliable predictions. Current model development sample size calculations are mainly conducted before data collection, leading to a fixed minimum sample size target based on sensible assumptions. However, adaptive sample size calculations can be used *during* data collection, to sequentially examine expected model performance and identify when enough data have been collected.

This study aims to extend existing sequential sample size calculations when developing a CPM, by applying stopping rules based on individual-level uncertainty of estimated risks and probability of misclassification. This is relevant for situations including prospective cohort studies with a short-term outcome.

Methods Using a sequential approach, the model development strategy is repeated after every 100 new participants are recruited, beginning when the initial sample size reaches the minimum recommended before analysis. For every iteration of the model, prediction and classification instability statistics and plots are calculated using bootstrapping, alongside measures of calibration, discrimination and clinical utility. For each statistic, learning curves display the trend of estimates against sample size and stopping rules are formed on the perceived value of additional information; crucially this is context specific, for example, guided by the level of uncertainty and classification errors that stakeholders (e.g., patients, clinicians) are willing to accept.

Results Our approach is illustrated using real examples, including (penalised and unpenalised) regression and machine learning approaches. The findings show that the sequential approach often leads to much larger sample sizes than the fixed sample size approach, and learning curves based on individual-level stability typically require larger sample sizes than focusing on population-level stability defined by overall calibration, discrimination, and clinical utility. Further, what ultimately constitutes an adequate sample size is strongly dependent on the level of prediction and classification instability deemed acceptable by stakeholders.

Conclusions For model development studies carrying out prospective data collection, an uncertainty-based sequential sample size approach allows users to dynamically monitor and identify when enough participants have been recruited to minimise prediction and classification instability in individuals. Engagement with patients and other stakeholders is crucial.

3: Determining the Sample Size for Risk Prediction Models with Clustered Binary Data

Izzati Izyani Japar¹, Gareth Ambler¹, M. Shafiqur Rahman², Rumana Omar¹

¹Statistical Science, University College London, United Kingdom

²Institute of Statistical Research and Training (ISRT), University of Dhaka, Bangladesh

Background Risk prediction models are increasingly being used in clinical practice to predict health outcomes. These models are often developed using data from multiple centres (clustered data) where patient outcomes within a centre are likely to be correlated. It is

important that the dataset used to develop a risk model is of an appropriate size, to avoid model overfitting problems and poor predictions in new data. Wynants *et al.* (2015) recommended using at least 10 events per variable (including the random parameter) to minimise bias in the regression coefficients and obtain acceptable C-statistic values when applying a random-effects model to clustered data. This approach focused only on ‘median predictions’ where the random intercept is ignored. More recently, Riley *et al.* (2020) and Pavlou *et al.* (2024) have proposed methods for sample size for independent data targeting the predictive performance of models however, these methods may not be appropriate for clustered data.

Methods We conducted a full-factorial simulation to evaluate whether the Wynants method yields sample sizes sufficient for developing models with good predictive ability. Additionally, we assessed the applicability of sample size methods proposed by Riley and Pavlou for clustered data. Simulation scenarios were investigated by varying multiple factors (e.g., degree of clustering, model strength etc.). Model performance was evaluated using the mean absolute prediction error (MAPE), calibration slope (CS), and the c-statistic. Both overall and cluster-specific performance measures were used and acceptable target values were specified for these measures. We developed a new sample size calculation formula for clustered data using a meta-model based on the simulation results.

Results None of the existing methods achieved our target acceptable MAPE values. The approaches by Wynants and Riley failed to achieve a CS of at least 0.9 when the prevalence was 15%. All methods generally produced c-statistic values within 0.02 of their true values. The new meta-model formula generally achieved the target acceptable MAPE values. It produced CS of at least 0.9 and c-statistic values within 0.02 of their true values when the prevalence was 25%.

Conclusions Current sample size calculation methods for developing binary risk models often fail to ensure adequate predictive performance of models and therefore may not be suitable for clustered data. A novel sample size calculation formula that achieved good predictive performance of the models across a range of clustered data scenarios is proposed.

4: Conformal Prediction Intervals for the Individual Treatment Effect

Danijel Kvaranovic², Robin Ristl¹, Martin Posch¹, Hannes Leeb²

¹Medical University of Vienna, Austria

²University of Vienna, Austria

Background The analysis of randomized clinical trials typically focusses on the estimation

of the average treatment effect. However, the effect of a medical treatment in a specific patient may depend on individual patient characteristics. Predictions regarding this individual treatment effect have the potential to allow for personalized treatment decisions and improve overall treatment success. To allow for an informed decision, prediction intervals are required, which take into account model uncertainty and individual residual variability, and cover the true individual treatment effect with a certain probability. In absence of strong model assumptions, the calculation of such intervals from parallel-group data is complicated by the fact that for each patient the outcome under only one treatment option can be observed and the counterfactual outcome remains unknown.

Methods We consider the setting of a randomized clinical trial comparing an experimental treatment versus control. We propose several procedures to calculate prediction intervals for the individual treatment effect in a new patient, which use multidimensional patient characteristics and a prediction model that is fitted with data from the randomized trial. The proposed methods do not depend on the chosen prediction model, however, for illustration, we consider linear regression models and fully connected neural networks. To construct the prediction intervals for the individual treatment effect, we first use two variations of the conformal inference method [Vovk, Gammerman and Shafer. Algorithmic learning in a random world. Springer, New York, 2005] to construct prediction intervals of the outcome under either treatment or control. In a second step we combine these intervals to obtain a prediction interval for the difference of the individual outcomes under treatment and control. In a simulation study, we compare the coverage probability and length of the proposed intervals using different regression models and under different assumptions regarding the distribution of individual residuals.

Results We analytically prove finite-sample coverage guarantee for two prediction interval procedures with mild assumptions on the true data generating process. We prove asymptotic coverage for a further method that allows for more narrow intervals, however requires a consistent regression model and bivariate normal distribution of the individual residuals. We further demonstrate that complex learning algorithms, such as neural networks, can lead to narrower prediction intervals than simple algorithms, such as linear regression, if the sample size is large enough.

Conclusions The proposed methods provide robust prediction intervals for the individual treatment effect, which have the potential to support personalized treatment decisions.

5: Effective Sample Size for Cox Survival Models: A Measure of Individual Uncertainty in Predictions

Toby Hackmann¹, Doranne Thomassen¹, Saskia le Cessie^{1,2}, Hein Putter¹, Liesbeth C de Wreede¹, Ewout W Steyerberg^{1,3}

¹Department of Biomedical Data Sciences, LUMC, The Netherlands

²Department of Clinical Epidemiology, Leiden University Medical Center, the Netherlands

³Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, the Netherlands

Background/Introduction Clinical prediction models are becoming increasingly popular to support shared decision-making. Most prediction models prioritize the accurate estimation and clear communication of point predictions. Uncertainty around the point prediction may be expressed by confidence intervals but is usually left out altogether. To present prediction uncertainty in an intuitive way, the concept of effective sample size may be attractive.[1] Our goal is to provide estimates of the effective sample size for individual survival predictions at a specific timepoint based on a Cox model.

Methods Effective sample size for a patient's risk prediction is defined as the hypothetical sample size of similar patients (with respect to the model), such that the variance of the survival probability in that sample would be the same as the prediction variance. In many cases, it can be calculated as a ratio of the outcome variance conditional on the predictor values to the prediction variance. We estimate the effective sample size for Cox model-based risk predictions using standard variance formulae. We investigate the behaviour of this estimator in an illustrative clinical data set of colon cancer patients and through simulations.

Results The variance of a risk prediction based on a Cox model depends on the variance of the estimated coefficients and variance of the baseline hazard. The latter is impacted by censoring and calculated based on the complete dataset weighted by the linear predictors. Effective sample size for a prediction is impacted by the distribution of covariates and censoring pattern in the development data and by model assumptions. Patients who are more 'typical'/well-represented in the data, with covariate values close to the population mean, have higher effective sample sizes, while patients with more uncommon covariate values have lower effective sample sizes. Model assumptions, such as the proportional hazards assumption and resulting shared baseline hazard in the Cox model, increase the effective sample size of predictions, sometimes to counterintuitively high values.

Conclusions Effective sample size can express the statistical/sampling uncertainty of risk predictions from a Cox model for individual patients. This uncertainty measure could be better interpretable for healthcare providers or patients compared to estimates of variance or

Abstracts of Contributed Talks

confidence intervals. Future studies should clarify its role in communicating uncertainty of predicted survival probabilities.

References [1] Thomassen, D., le Cessie, S., van Houwelingen, H. C., & Steyerberg, E. W. (2024). Effective sample size: A measure of individual uncertainty in predictions. *Statistics in Medicine*. <https://doi.org/10.1002/sim.10018>

Defining Estimands for Clinical Trials

Monday, 2025-08-25 14:00 - 15:30, Biozentrum U1.131

Chair: Kelly Van Lancker

1: What is the Estimand for the Proportional Odds Model?

Ian R White, Brennan C Kahan

UCL, United Kingdom

Background The proportional odds model is frequently used to analyse an ordered categorical outcome in clinical trials. It assumes that the odds ratio is the same for all dichotomisations of the outcome.

We wanted to know what estimand the model estimates if this proportional odds assumption is false. The log odds ratio from a proportional odds model is known to be a function of the probability that a randomly selected person from the treatment group has a better outcome than a randomly selected person from the control group. However, this is hard to interpret. To interpret the log odds ratio from a proportional odds model better, we explore whether it can be expressed as a difference of an average score between the two groups (where the score is some transformation of the outcome).

Methods We propose two novel approximate approaches which demonstrate how the proportional odds model contrasts the average scores between groups. Mathematically, we derive a closed-form expression for the maximum likelihood estimator for data near the null. Computationally, we estimate scores by equating influence (change in the log odds ratio due to omitting one observation) between the proportional odds model and a linear regression for scores. The methods are illustrated using the results of the FLU-IVIG trial.

Results The mathematical derivation shows that the log odds ratio can be approximated as a difference of mean scores between groups, where the score for each outcome level is a linear function of the proportion below that level plus half the proportion at that level. The computational method agrees well with the mathematical expression, and can also be used away from the null or with covariate adjustment.

The FLU-IVIG trial had rare categories representing very poor clinical outcomes. The derived scores implied by the proportional odds model are shown to be very similar for these worst categories. This may be undesirable, since they have very different clinical importance.

Conclusion The estimand for the proportional odds model is a difference of mean scores which are implicitly assigned by the model. Adjacent small categories have very similar implicit scores. Deriving scores provides a way to discuss the proportional odds model with clinicians, who can thus decide whether it suitably reflects clinical importance.

2: On the Use of the Net Treatment Benefit as a Treatment-Effect Measure in Randomized Clinical Trials

Tomasz Burzykowski^{1,2}, Vaiva Deltuvaitė-Thomas²

¹Hasselt University, Belgium

²IDDI, Belgium

Generalized pairwise comparisons (GPC) is a non-parametric method designed to quantify the benefit of a new treatment, as compared to a control one, by using a set of hierarchically-ordered endpoints. GPC yields an estimate of the treatment effect, the so-called net treatment benefit (NTB). For a single endpoint, $NTB(d) = P(X^E - X^C > d) - P(X^E - X^C < -d)$, where X^E and X^C is the value of the endpoint for the experimental and control group, and d is the threshold of clinical relevance. In this case, $NTB(d)$ is simply the difference between the probability that the value of the endpoint for the experimental treatment is clinically “better” than for the control treatment and the probability that the value of the endpoint for the control treatment is “better” than for the experimental treatment.

GPC and NTB are being promoted as the approach that allows benefit-risk assessment and that is useful for personalized medicine. In this paper, we take a critical look at some of the properties of NTB that may be important from a point of view of using it as a treatment-effect measure in randomized clinical trials. Several of the properties have been already investigated in the literature. For instance, NTB may be trial-specific, because it depends on the variability of endpoint(s) which may change from trial to trial. NTB is not robust to missing data, and its application to right-censored data requires the use of corrections. We additionally show that, in case hidden strata are present in a population, they may lead to biased estimation of NTB in a clinical trial. For instance, it is possible to obtain a non-zero NTB value when the true value is zero. We also show that NTB may suggest no benefit of either treatment when, in fact, the patients receiving an experimental treatment fare, on average, worse on it (as compared to patients receiving the control treatment) when the treatment does not work than when the treatment does work.

3: Challenges and Opportunities in Defining New Estimands for Longitudinal Outcomes Truncated by Death

Juliette Ortholand¹, Marie-Abèle Bind²

¹Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, F-75013, Paris, France

²MGH Biostatistics, Massachusetts General Hospital, Harvard Medical School, Boston

Introduction Quality of life, often measured through repeated measures in clinical trials, is a meaningful outcome for patients, even for deadly diseases. Nevertheless, studying longitudinal outcomes that are truncated by death raises methodological challenges in defining suitable estimands. Indeed, when patients die during a clinical trial, their longitudinal outcomes are not defined at the end of the clinical trial. Although a few estimands have been proposed to address this issue, many lack a formal causal interpretation.

Objective In our work, in line with guidelines of the European Medicines Agency, we discuss the relative advantages of different causal estimands for longitudinal outcomes truncated by death.

Method First, to better understand the challenges of estimands in this context, we start with a set-up without truncation. Then, we compare the different estimands proposed in the literature from a causal point of view and discuss their estimation. Finally, we illustrate this work using clinical trial data from amyotrophic lateral sclerosis (ALS).

Results We find two primary challenges in defining estimands for longitudinal outcomes truncated by death. First, defining an order of better health in this situation is not straightforward. Second, when the average causal effect is used as population-level causal effects, a notion of distance between dead and alive individuals must be defined, in addition to an order of better health.

We show that principal stratification allows us to avoid the need for an ordering between dead and alive individuals by focusing on an "always survivor" subgroup. Furthermore, when assuming death as the worst outcome (for the sake of defining an order), we present the causal framework of other estimands, such as combined scores and the survival-incorporated median, which rely on well-selected population-level causal effects to circumvent the issue of defining a distance between dead and alive individuals.

Our work also shows how these estimands target different scientific questions and highlights how the use of estimators can help alleviating some assumptions necessary for estimation.

Conclusion This work offers insights into better incorporating quality of life as an endpoint in clinical trials and provides a clearer framework for utilizing powerful estimators in the challenging context of longitudinal outcomes truncated by death.

4: What Estimands can and Should be Used when Interventions may be Given more than Once?

Joanna Anetta Hindley¹, Michelle Clements¹, James Carpenter^{1,2}, Brennan Kahan¹

¹University College London, United Kingdom

²London School of Hygiene and Tropical Medicine, United Kingdom

Introduction In many settings, patients may require treatment more than once. For instance, patients who experience asthma exacerbations might require medication at each onset of symptoms, and those undergoing IVF may receive several rounds of treatment until they become pregnant. Historically, most randomised trials only allowed participants to enrol for a single ‘treatment episode’, but there is growing recognition that allowing participants to re-enrol for each new episode they experience can be beneficial. However, doing so poses additional challenges around defining the estimand, as investigators need to consider issues such as how different patients experience different numbers of episodes, or the fact that patients may be assigned different treatments at different episodes. To date, there has been little work on categorising the variety of estimands that can be used in these settings, comparing the interpretability of each, or evaluating statistical estimators.

Methods/Approach We define a range of estimands that can be used in these ‘multi-episode’ settings. We discuss the interpretation of each and their clinical relevance in settings like IVF or asthma exacerbations, then evaluate the performance of different trial designs (re-randomisation and single-patient cluster randomisation) and statistical estimators for each using a simulation study. Within this simulation we allow for different data-generating models for both outcome measure and episode occurrence.

Results We define 8 estimands suitable for settings where treatments may be given more than once. Of these, 6 can be estimated under minimal assumptions so long as an appropriate design and estimator are used; however, the remaining 2 estimands require strong, untestable assumptions, and are therefore unlikely to be useful in a primary analysis. By considering the contexts of IVF and asthma exacerbations we find that the relevance of each estimand may vary depending on clinical context, highlighting the importance of careful consideration of this issue at the trial design stage.

Discussion We describe estimands suitable for use in trials evaluating treatments that may be given more than once, show that most can be estimated under minimal assumptions and explore considerations for choosing the most useful estimand. This work brings clarity to the design and analysis of clinical trials in settings that are common across medical specialties.

5: Current Practice Around the Use of Estimands in Cluster Randomised Trials, and the Impact of Informative Cluster Size on Inferences

Dongquan Bi, Andrew Copas, Brennan Kahan

Institute of Clinical Trials and Methodology, University College London, United Kingdom

Introduction The use of estimands helps clarify the treatment effect a study aims to estimate. Cluster randomised trials (CRTs) can address the participant-average (PA) or cluster-average (CA) effects. When outcomes and/or treatment effects vary across clusters depending on cluster size (also called ‘informative cluster size’ (ICS)), these two effects can differ, and estimators that target one can be biased for the other. It has been recently shown that common estimators for CRTs such as generalised estimating equations with an exchangeable correlation structure (GEE) and mixed-effects models can produce biased estimates for both PA and CA effects under ICS. However, current practice around the use of estimands in CRTs and the likely impacts of ICS remain unknown. We therefore set out to establish the current practice through a review and explore the potential impacts of ICS in a re-analysis of a CRT.

Methods We conducted a review of recently published CRTs to explore which estimands are targeted, which estimators are used, and how often potential impacts of ICS are considered. We then reanalysed the RESTORE trial, which randomised 31 US paediatric intensive care units to compare protocolised sedation with usual care for critically ill children. For each outcome, we compared the PA and CA effect estimates from independence estimating equations (IEE) which is robust to ICS, to evaluate the likelihood of ICS. Next, we compared estimates from GEE/mixed-effect models against IEE estimates to evaluate the extent to which results from GEE and mixed-effects models may have been affected by ICS. We used bootstrapping to evaluate how much the difference could be due to chance.

Results No trial (0/73) tried to report estimands. The research question was not clear in most trials (58/73). For many trials (46/73), it was not inferable whether they were targeting a PA or CA estimand. Trials often used GEE or mixed models as the primary estimator (37/73). The potential impacts of ICS were rarely considered.

The re-analysis found that the PA and CA estimates differed for most outcomes (18/22), indicating a possible presence of ICS in the RESTORE trial. For instance, for the iatrogenic withdrawal outcome, the PA OR estimate was 1.35 (95% CI, 0.72 - 2.51), and the CA OR estimate was 0.81 (95% CI, 0.39 – 1.69).

Potential impact Trialists should describe estimands in CRTs which helps to ensure that research question investigated can be understood and appropriate statistical methods aligned with the question are used.

Adaptive and Multi-Arm Multi-Stage Trials

Monday, 2025-08-25 14:00 - 15:30, Biozentrum U1.141

Chair: Nigel Stallard

1: Confidence Intervals in Two-Stage Adaptive Enrichment Designs

Enyu Li¹, Nigel Stallard¹, Ekkehard Glimm², Dominic Magirr², Peter Kimani¹

¹University of Warwick, Coventry, United Kingdom

²Novartis Pharma AG, Basel, Switzerland

Background With the deepening understanding of biological pathways of disease progression, the analysis of subpopulations has gained importance in clinical trials. Two-stage adaptive enrichment designs have been proposed as an efficient approach for subgroup analysis when treatment heterogeneity is suspected. In such a design, in stage 1, patients are recruited from the full population. Then, the subpopulation that appears to benefit from the experimental treatment is selected by an interim analysis based on stage 1 outcomes data. In stage 2, patients are only recruited from the selected population. Data from both stages are used in the final confirmatory analysis. The selection nature of adaptive enrichment designs poses statistical challenges regarding inference for treatment effects. In this work, we develop selection adjusted confidence intervals for adaptive enrichment designs.

Method We consider a two-stage adaptive enrichment design based on Jenkins et al [1]. Based on statistical theories in Lehmann and Romano [2], conditional on each possible decision at the interim analysis, we derive the uniformly most accurate unbiased confidence intervals by inverting the two-sided uniformly most powerful unbiased test and provide numerical methods for calculating the intervals.

Result As implied by the theoretical derivation, our method has the advertised coverage probability conditional on each possible interim decision. Through an intensive simulation study, we verified that confidence intervals using our method have coverage probabilities that are approximately equal to the nominal level. In contrast, previously proposed approaches including double bootstrap confidence intervals [3] and duality confidence intervals [4] show substantial under-coverage in some cases.

Reference [1] Jenkins, M., Stone, A. and Jennison, C. (2011), An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceut. Statist.*, 10: 347-356. <https://doi.org/10.1002/pst.472>

- [2] Lehmann, E.L. and Romano, J.P. (2022). Testing statistical hypotheses. New York: Springer. <https://doi.org/10.1007/978-3-030-70578-7>
- [3] Magnusson, B.P. and Turnbull, B.W. (2013), Group sequential enrichment design incorporating subgroup selection. *Statist. Med.*, 32: 2695-2714. <https://doi.org/10.1002/sim.5738>
- [4] Kimani PK, Todd S, Renfro LA, et al. Point and interval estimation in two-stage adaptive designs with time to event data and biomarker-driven subpopulation selection. *Statistics in Medicine*. 2020; 39: 2568–2586. <https://doi.org/10.1002/sim.8557>

2: Dynamic Bayesian Sample Size Re-Estimation: Balancing Historical and Real-Time Data in Adaptive Clinical Trials

Niamh Anne Fitzgerald¹, James Wason¹, Adrian Mander²

¹Newcastle University, United Kingdom

²GSK

Background Sample size re-estimation (SSRE) is a key component of adaptive clinical trials, allowing for mid-trial adjustments to improve efficiency and statistical power. Traditional frequentist SSRE methods rely on interim variance estimates, which may be imprecise, particularly in small sample sizes. Bayesian methods offer an alternative approach by incorporating external historical data, which can improve variance estimation and lead to better-informed sample size adjustments. However, excessive reliance on historical data may introduce errors if past studies do not align well with the current trial. This study explores a Bayesian SSRE method that dynamically balances historical and real-time data to improve sample size estimation.

Methods We conducted Monte Carlo simulations to evaluate Bayesian and frequentist SSRE strategies in a two-arm clinical trial setting. The frequentist approach re-estimates sample sizes based on pooled variance from interim analyses, while the Bayesian approach integrates historical data using a mixture prior that shifts weight from external information to accumulating trial data. Each method was assessed based on how closely the re-estimated sample size aligned with the sample size that would be chosen if the true parameters were known from the outset.

Results Our findings show that Bayesian SSRE methods provide more accurate sample size adjustments compared to frequentist methods. The dynamic mixture prior approach effectively transitions reliance from external data to real-time trial information, leading to

more precise sample size determinations. The Bayesian method consistently produced sample sizes that were closer to the ideal sample size, particularly when the assumed variance differed from the true variance. Additionally, we examined the effect of different historical data weights, demonstrating that moderate incorporation of external data improves estimation accuracy, while excessive reliance on historical information can lead to deviations.

Conclusion Bayesian SSRE methods offer notable advantages for adaptive clinical trials by improving sample size estimation and reducing the risk of under- or over-enrolment. By dynamically adjusting the contribution of historical data, these methods enhance trial efficiency and reliability while maintaining statistical rigour. Our findings support the adoption of Bayesian SSRE approaches in modern clinical trial designs, particularly in settings where variance assumptions are uncertain.

3: MAMS with Patient Reported Outcomes and Treatments with Different Modalities: How Should We Handle a Differential Placebo Effect?

Isobel Grace Landray¹, Jennifer Nicholas¹, James Carpenter^{1,2,3}

¹London School of Hygiene and Tropical Medicine, United Kingdom

²MRC CTU at UCL, United Kingdom

³On behalf of the Edmond J Safra Accelerating Clinical Trials in Parkinson's collaboration

Background MAMS-platform trials have demonstrated critical efficiency improvements in identifying effective new treatments. The design is now well-accepted, leading to continuing increases in diversity and complexity of applications.

A key example of this is the individually-randomised EJS ACT-PD(23) MAMS-platform, which will open in Summer 2025 with two active arms and a placebo, all delivered through a daily blinded capsule. The primary outcome is a subjective Parkinson's specific patient-reported functional score. A fourth arm, to be added in Summer 2026, will comprise five daily capsules.

The resulting methodological challenge is: how to identify efficient and resilient design solutions which maintain the integrity of all treatment comparisons while minimising control group size when treatment regimens differ.

Methods When the new arm is added, there are three main options:

1. keep the size of the placebo arm the same, and further randomise the placebo patients

- 1:1 to one or five capsules daily;
2. add a second placebo arm which takes five capsules daily, and maintain this till the end of the trial, or
 3. add a second placebo arm which takes five capsules daily, but after one year, test for a difference between placebo groups; if there is no difference, move to option 1.

We will use theoretical and simulation based methods to evaluate the consequences of increasing levels of a differential placebo effect on the power of the study.

Results Option 2 is the gold standard which preserves the integrity of all treatment comparisons. However, because it increases the probability of randomisation to a placebo, it is less acceptable to trial participants.

We will present graphs of power versus effect size for different levels of differential placebo effect, and show how these can be used to identify the bound on a differential placebo effect which is sufficient to justify option 1 and/or inform the test in option 3.

Conclusion A key attraction of MAMS studies is the option to add new arms as new treatments emerge. Our results show how to frame the discussion on how to adapt the design in settings where the outcome is potentially subjective, so participants must be blinded, and the new treatments have a range of regimens.

4: Multi-Arm Multi-Stage (MAMS) Randomised Selection Designs: With an Application to Miscarriage and Surgical Platform Trials

Babak Choodari-Oskooei¹, Alexandra Blenkinsop², Lee Middleton³, Kelly Handley³, Versha Cheed³, Lee Priest³, Emily Fox³, Leah Fitzsimmons³, Rima Smith³, Adam Devall³, Thomas Pinkney³, Arri Coomarasamy³, Mahesh KB Parmar¹

¹UCL's Institute of Clinical Trials and Methodology, United Kingdom

²Department of Mathematics, Imperial College London, United Kingdom

³Birmingham University, United Kingdom

Background Multi-arm multi-stage (MAMS) randomised trial designs have been proposed to evaluate multiple research questions in confirmatory settings. In designs with several interventions, there are likely to be strict limits on the number of individuals that can be recruited or the funds available to support the protocol. These limitations may mean that

not all research treatments can continue to accrue the required sample size for the definitive analysis of the primary outcome measure at the final stage. In these cases, an additional treatment selection rule can be applied at the early stages of the trial to restrict the maximum number of research arms that can progress to the subsequent stage(s).

This talk provides guidelines on how to implement treatment selection within the MAMS framework. It explores the impact of treatment selection rules, interim lack-of-benefit stopping boundaries, the timing of treatment selection as well as using an intermediate outcome to select interventions on the operating characteristics of the design. It uses real trials such as Tommy's PREMIS (miscarriage) and ROSSINI-2 (surgery, 8-arm 3-stage) to compare MAMS selection designs with alternative designs.

Methods We outline the steps to design a MAMS selection trial. Using extensive simulation studies, we explore the maximum/expected sample sizes, familywise type I error rate (FWER), and overall power of the design under binding and non-binding interim stopping boundaries for lack-of-benefit.

Results Pre-specification of a treatment selection rule reduces the maximum sample size by at least 25%. The familywise type I error rate of a MAMS selection design is smaller than that of the standard MAMS design with similar design specifications without the additional treatment selection rule. In designs with strict selection rules, the final stage significance levels can be relaxed for the primary analyses to ensure that the overall type I error for the trial is not underspent. When conducting treatment selection from several treatment arms, it is important to select a large enough subset of research arms (that is, more than one research arm) at early stages to maintain the overall power at the pre-specified level.

Conclusions MAMS selection designs gain efficiency over the standard MAMS design by reducing the overall sample size. Diligent pre-specification of treatment selection rules, final stage significance level and interim stopping boundaries for lack-of-benefit are key to controlling the operating characteristics of a MAMS selection design. We provide guidance on these design features to ensure control of the operating characteristics.

5: A Test for Treatment Differences using Allocation Probabilities in Response-Adaptive Clinical Trials

Stina Zetterstrom, David S. Robertson, Sofía S. Villar

MRC Biostatistics Unit, University of Cambridge, United Kingdom

Background/Introduction:

Response-adaptive clinical trials update the allocation probabilities to treatment arms sequentially based on the data collected so far in the trial. One main objective for response-adaptive clinical trials is to ensure that patients in the trial will have a higher probability of getting the best treatment compared to using equal randomisation. However, this imbalance in treatment allocation can lead to low power when testing for a treatment difference. Recent works (Barnett et al 2024; Deliu et al 2023) propose a new testing approach that is based on the allocation probability (AP) instead of the outcome directly, which can increase the power. Here, we further investigate and propose alternative versions of the AP test.

Methods:

The original AP test statistic is the sum of blocks in the trial with an allocation probability, to the active arm, larger than 0.5 (in a two-arm trial). The null and alternative distributions of the test statistic are found using simulations. We propose alternative versions of the AP test, where the magnitude of the allocation probabilities and/or the block number is utilised in the test statistic. The different versions of the AP test are evaluated in simulation studies, using the Thompson sampling algorithm for binary outcome in a two-arm setting.

Results:

The simulation studies show that the AP test can perform better in terms of power compared to traditional tests, especially in certain regions of the parametric space, while controlling type 1-error. Furthermore, the AP tests that uses the magnitude of the allocation probabilities and/or the block number can have a higher power than the original AP test, especially when the trial size is larger.

Conclusions:

The AP test is an alternative hypothesis test that can be used in response-adaptive clinical trials when the treatment allocation is imbalanced to increase the power compared to traditional methods. There are different versions of the test, that can further increase the power gain, in specific settings. While we use Thompson sampling, the AP test can be used for any response-adaptive randomisation, and its performance studied in that intersection.

References:

Helen Yvette Barnett, Sofía S. Villar, Helena Geys, Thomas Jaki, "A Novel Statistical Test for Treatment Differences in Clinical Trials Using a Response-Adaptive Forward-Looking Gittins Index Rule", *Biometrics*, 79(1), 2023, Pages 86–97.

Abstracts of Contributed Talks

Nina Deliu, Joseph J. Williams, and Sofía S. Villar. "Efficient inference without trading-off regret in bandits: An allocation probability test for Thompson sampling." *arXiv preprint arXiv:2111.00137*, 2021.

Causal Inference: Target Trial Emulation

Monday, 2025-08-25 14:00 - 15:30, Biozentrum U1.101

Chair: Shaun Seaman

1: Inference on Sustained Treatment Strategies, with a Case Study on Young Women with Breast Cancer

Elise Dumas^{1,2}, Floriane Jochum², Florence Coussy², Anne-Sophie Hamy², Sophie Houzard³, Christine Le Bihan-Benjamin³, Fabien Reyal², Paul Gougis², Mats Julius Stensrud¹

¹Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

²Institut Curie, France

³Institut National du Cancer, France

Background Lack of adherence can reduce the effectiveness of beneficial treatments. However, the extent to which adherence affects outcomes is unclear in many settings. For example, is it enough to adhere to treatment for 80% or 90% of the prescribed days? Does it matter whether adherence is higher earlier or later in the treatment schedule? And is it particularly important not to miss consecutive days of treatment? In this work, we explore a methodology to emulate and compare different adherence strategies. Specifically, we use a framework that incorporates explicit grace periods and regimes that depend on natural treatment patterns. Our work is motivated by a clinical question concerning women with early-stage hormone receptor-positive breast cancer, for whom daily endocrine therapy is prescribed for five to ten years. In these patients, young age is associated with both an increased risk of cancer recurrence and suboptimal adherence to endocrine therapy.

Methods Using French nationwide claims data, we applied the proposed methods to compare the survival benefits achievable in patients with breast cancer by implementing different adherence strategies to endocrine therapy for each age group. We emulated three different adherence strategies that allowed for gaps in treatment of no more than one, three, or six consecutive months.

Results A total of 121,601 patients were included in the analyses. In patients aged 34 years or younger, strict ET adherence (1-month gaps) improved 5-year DFS by 4.3 percentage-points, (95% confidence interval (CI): 2.6-7.2) compared to observed adherence. In this age group, ET adherence strategies allowing for 3-month and 6-month gaps reduced the 5-year DFS benefit to 1.3 (95% CI: 0.2-3.7) and 1.0 (95% CI : -0.2-3.4) percentage-points,

respectively. In contrast, DFS benefits of improved ET persistence in patients after 50 years old did not exceed 1.8 percentage-points, compared to observed persistence, regardless of the length of gaps allowed.

Conclusion Our results show that young women would benefit substantially from stricter adherence to endocrine therapy, with treatment breaks never exceeding one month, highlighting the need for tailored strategies to improve treatment adherence in this population.

2: An Overlooked Bias in Target Trial Emulations and How to Fix it

Lorenzo Gasparollo, Mats Stensrud

EPFL

Many datasets involve staggered entries, where individuals join the study at different points in time. For example, randomized controlled trials (RCTs) in medicine usually recruit patients over time, and electronic health records contain information from the time a patient enters the healthcare system. Seminal works by Hernán, Brumback, and Robins (2000) and Hernan et al. (2008) on target trials used such datasets, treating the time an individual entered the study as a covariate in regression models. In this talk, I will describe a subtle - but frequently overlooked - positivity violation that appears in the analysis of staggered entry data. Because of this positivity violation, a frequently used class of Inverse Probability Weighting Censoring procedures leads to biased results. I will then propose new adjustment methods to circumvent such bias, and elaborate on how these fit with the target trial emulation framework. Finally, I will outline and compare these two approaches in settings wherein the bias varies in severity, thereby clarifying when one method is preferred over the other.

3: Model-Free Estimands for Target Trial Analysis

Edoardo Efrem Gervasoni, Oliver Dukes, Stijn Vansteelandt

Ghent University, Belgium

The target trial framework is a powerful methodology to estimate causal effects in observational studies by emulating randomized controlled trials. It addresses common biases in observational data, such as confounding and selection bias, by conceptualizing a sequence

of hypothetical trials initiated at different time points. To increase precision, information is pooled across trials, typically under the assumption that the treatment effect is constant over time and across individuals. However, this can create ambiguity about the causal question when the effects vary or, as frequently happens, the population observed in some emulated trials systematically differs from the target population. Further challenges come when effects are parametrized using non-linear regression models (e.g. logistic regression) where non-collapsibility can create issues of misspecification when pooling different populations. To address these challenges, this project introduces an assumption-lean strategy for target trial analysis, focussing on the choice of the estimand, rather than the choice of a model. This ensures that the analysis' aim is unequivocal regardless of model misspecification, and that uncertainty assessments reflect only information available in the data. Our proposal consists of several estimands, each related to different data structures and addressing different aspects of the patient population that may be of interest to researchers or decision-makers. For these estimands, corresponding estimators have been developed by the use of G-computation and inverse probability weighting. Applications on simulations and real data on antimicrobial de-escalation in an intensive care unit setting demonstrate the advantages of the proposed methodology over traditional techniques, offering greater clarity and reliability in causal effect estimation.

4: Target Trial Emulation: Optimising Methods for Estimating Treatment Effects using Data from the UK Cystic Fibrosis Registry

Emily Granger¹, Lorna Allen², Susan Charman², Sarah Clarke², Gwyneth Davies³, Freddy Frost⁴, Laurie Tomlinson¹, Ruth Keogh¹

¹London School of Hygiene and Tropical Medicine, United Kingdom

²Cystic Fibrosis Trust

³UCL Great Ormond Street Institute of Child Health

⁴University of Liverpool

Background A key challenge in medical statistics is that there remain many questions about the effects of treatments that are unlikely to be addressed in randomised trials. For example, in cystic fibrosis, long-term treatment effectiveness is a priority research area, but is difficult to address in trials due to feasibility and cost. When a randomised trial is not feasible, an alternative is to use observational data to 'emulate' a target trial. Target trial emulation (TTE) helps to avoid commonly occurring biases in observational studies, but it is unclear how best to apply TTE to data from existing cystic fibrosis registries.

The UK Cystic Fibrosis (CF) Registry, managed by the CF Trust, collects data on over 99%

of the UK CF population. It includes longitudinal data on clinical variables collected at annual review visits and data on treatment prescription start and stop dates. Our aim is to establish best practices for the application of TTE using UK CF Registry data to estimate the effects of treatments in CF.

Methods and Results We describe different methodological approaches to applying TTE to these data; a key consideration is how to define time 0. Previous studies applying TTE using these data have defined time 0 as the date of the annual review in which individuals meet the eligibility criteria for the emulated trial. They assume that treated individuals began treatment on the date of the annual review. In reality, individuals may have initiated treatment at any point between two consecutive annual review visits. An alternative approach is to make use of the data on prescription dates and there are different ways that time 0 could be defined for treatment initiators and non-initiators using these data. Another challenge is that the outcomes are often measured at different times (relative to time 0) for different individuals. The proposed methodological approaches are compared in a series of trial emulations of published trials in cystic fibrosis.

Conclusions TTE is a commonly used approach to investigating treatment effectiveness using observational data, when a randomised trial is not feasible. However, it is not always clear how best to implement TTE using the available data. A key outcome of our study will be practical guidance for the application of TTE using UK CF Registry data. Our findings may be useful for other disease areas that benefit from patient registries.

5: Target Trials and Structural Nested Models: Emulating RCTs using Observational Longitudinal Data

Oliver Dukes¹, Fuyu Guo², Mats Stensrud³, James Robins²

¹Ghent University (Belgium)

²Harvard School of Public Health (USA)

³EPFL (Lausanne)

Target trial emulation is a popular method for estimating effects of treatment regimes from observational data. In the emulation, new trials, indexed by time, are initiated at fixed intervals. A subject participates in every trial for which eligibility criteria are met. Current methods treat each time-specific trial separately. For instance, for a trial comparing the regimes “always” versus “never” treat from initiation at t onwards, it is common to fit a hazard or risk ratio (RR) model that includes a treatment indicator and its potential confounders. Subjects are censored if they later change treatment, with inverse probability

Abstracts of Contributed Talks

weighting to adjust for the censoring. If most subjects change treatment, the estimates will be inefficient. In this paper we propose more efficient estimators by introducing regime-specific structural nested target trial emulation models (SNTTEM). Given a regime, a SNTTEM imposes parametric models for all time-specific blip functions of the eligible subjects and leaves those for the ineligible unrestricted. A time-specific blip function quantifies on a mean scale the effect of initiating the regime at a time t versus one period later, as a function of past history. The intersection of all the earlier time-specific RR models constitutes a single SNTTEM with regime “always take the treatment that one took last time”. We show that SNTTEM can be fitted using g-estimation, a method that censors less and is more efficient than current methods.

Meta-Analysis 1

Monday, 2025-08-25 14:00 - 15:30, ETH E27

Chair: Antonio Remiro-Azocar

1: Precision of Treatment Hierarchy: A Metric for Quantifying Certainty in Treatment Hierarchies from Network Meta-Analysis

Augustine Wigle¹, Audrey Bélieau¹, Georgia Salanti², Gerta Rücker³, Guido Schwarzer³, Dimitris Mavridis⁴, Adriani Nikolakopoulou^{5,3}

¹Department of Statistics and Actuarial Science, University of Waterloo, Ontario, Canada

²Institute of Social and Preventative Medicine, University of Bern, Bern, Switzerland

³Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany

⁴Department of Primary Education, University of Ioannina, Ioannina, Greece

⁵Laboratory of Hygiene, Social and Preventive Medicine and Medical Statistics, School of Medicine, Aristotle University of Thessaloniki, Thessaloniki, Greece

Background Network meta-analysis (NMA) is an extension of pairwise meta-analysis which facilitates the estimation of relative effects for multiple competing treatments. A hierarchy of treatments is a useful output of an NMA. Treatment hierarchies are produced using ranking metrics. Common ranking metrics include the Surface Under the Cumulative RANKing curve (SUCRA) and P-scores, which are the frequentist analogue to SUCRAs. Both metrics consider the size and uncertainty of the estimated treatment effects, with larger values indicating a more preferred treatment. Although SUCRAs and P-scores themselves consider uncertainty, treatment hierarchies produced by these ranking metrics are typically reported without a measure of certainty, which might be misleading to practitioners.

Methods We propose a new metric, Precision of Treatment Hierarchy (POTH), which quantifies the certainty in producing a treatment hierarchy from SUCRAs or P-scores. The metric connects three statistical quantities: The variance of the SUCRA values, the variance of the mean rank of each treatment, and the average variance of the distribution of individual ranks for each treatment. We show how the metric can be adapted to apply to subsets of treatments in a network, for example, to quantify the certainty in the hierarchy of the top three treatments.

Results We calculate POTH for a database of NMAs to investigate its empirical properties, and we demonstrate its use on two published networks: a network of antifungal treatments

to prevent mortality for solid organ transplant recipients ($\text{POTH}=0.326$) and a network of pharmacological treatments for persistent depressive disorder ($\text{POTH}=0.559$).

Conclusion Although POTH was proposed specifically to summarise the certainty in treatment hierarchies derived using SUCRAs or P-scores, it can also be viewed simply as a way to summarise all the ranking probabilities in a given network. It is therefore an indicator of the certainty in any treatment hierarchy derived using a ranking metric related to the ranking probabilities. In summary, POTH provides a single, interpretable value which quantifies the degree of certainty in producing a treatment hierarchy.

2: Extending P-Scores for Ranking Diagnostic Tests in Network Meta-Analysis

Sofia Tsokani^{1,2}, Fani Apostolidou-Kiouti¹, Adriani Nikolakopoulou^{1,3}, Areti-Angeliki Veroniki^{4,5}, Anna-Bettina Haidich¹, Dimitris Mavridis⁶

¹Laboratory of Hygiene, Social & Preventive Medicine and Medical Statistics, School of Medicine, Aristotle University of Thessaloniki, Thessaloniki, Greece

²Methods Support Unit, Cochrane, London, UK

³Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany

⁴Li Ka Shing Knowledge Institute, St. Michael's Hospital, Unity Health Toronto, Toronto, ON, Canada

⁵Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, ON, Canada

⁶Department of Primary Education, School of Education, University of Ioannina, Ioannina, Greece

Background / Introduction Network meta-analysis (NMA) of diagnostic test accuracy (DTA) studies allows for the comparison of multiple diagnostic tests in a single framework, considering both sensitivity and specificity. Traditional ranking metrics such as the diagnostic odds ratio (DOR) and superiority index summarize diagnostic performance but fail to account for the correlation between sensitivity and specificity. P-scores, initially suggested for ranking interventions in NMA, provide a probabilistic measure of how likely an intervention is superior to any other, averaged over all competing interventions. In NMA of interventions, there has been an extension of P-scores to accommodate multiple outcomes. Using this advancement, in this study, we adapt P-scores to DTA-NMA, incorporating sensitivity and specificity jointly to improve diagnostic test ranking.

Methods Our approach requires estimates of comparisons of sensitivity and specificity across

all diagnostic tests within the network, as derived from a DTA-NMA model. We use logit-transformed estimates of sensitivity and specificity, along with their standard errors and correlation between sensitivity and specificity. We have extended P-score methodology in which ranking probabilities for each test are computed by assessing test's superiority in both sensitivity and specificity simultaneously. To achieve this, we modified the P-score function, allowing it to accommodate bivariate diagnostic measures. The final rankings obtained via P-score were compared against DOR-based and superiority index-based rankings to assess differences in interpretation. An R package to enable implementation of P-scores in DTA-NMAs is under preparation.

Results As an example, we applied an ANOVA-based DTA-NMA model to a dataset of six diagnostic tests for breast cancer recurrence (15 studies, 659 individuals, 338 with recurrence, reference standard histological diagnosis/long-term clinical follow-up/autopsy finding). P-score rankings identified PET/CT as the best-performing test (P-score = 0.76), followed by MRI (0.58), PET (0.51), BS (0.39), CT (0.11), and CW (0.01). This means that PET/CT has 75% probability of outperforming all other diagnostic tests in both sensitivity and specificity. These results aligned with DOR rankings, where PET/CT had the highest DOR (88.99) and CW the lowest (7.00). However, P-scores provided additional insights, accounting also for the correlation between sensitivity and specificity, which traditional ranking methods overlook.

Conclusion Ranking diagnostic tests in DTA-NMA is challenging due to the bivariate nature of sensitivity and specificity. Traditional ranking methods fail to distinguish between tests with high sensitivity but low specificity (or vice versa). The proposed P-score extension offers a clearer, probabilistic ranking framework for more effective comparisons.

3: Resolving Conflicting Treatment Hierarchies Across Multiple Outcomes in Multivariate Network Meta-Analysis

Theodoros Evrenoglou¹, Anna Chaimani², Guido Schwarzer¹

¹Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center-University of Freiburg, Germany

²Oslo Center for Biostatistics and Epidemiology, Department of Biostatistics, University of Oslo, Oslo, Norway

Background/Introduction Multivariate network meta-analysis (mvNMA) extends network meta-analysis (NMA) by enabling the simultaneous comparison of multiple treatments across multiple outcomes. On top of indirect evidence, mvNMA accounts for the between-outcome

correlation, further improving the precision and reliability of the summary treatment effect estimates. Outputs from mvNMA are typically extensive and comprise several treatment effect estimates with varying uncertainty. Similar to NMA, these outputs can be summarized using ranking metrics. However, in mvNMA, separate treatment hierarchies are generated for each outcome, making it challenging to identify the best treatments overall, particularly when hierarchies conflict. To date, the literature lacks proper methods to address conflicting treatment hierarchies. Consequently, decision-making relies solely on ad-hoc approaches, based on separate NMAs for each outcome.

Methods We introduce a novel framework for handling conflicting treatment hierarchies across outcomes in mvNMA. First, we fit a Bayesian mvNMA model to obtain outcome-specific hierarchies in terms of SUCRAs. Then, to resolve conflicts across treatment hierarchies, we adapt the VIKOR method – originally proposed for multi-criteria decision analysis – to the meta-analytic setting. This method aims to identify the best ‘compromise’ solution across the different outcome-specific hierarchies. Specifically, by combining both the weighted L_1 and L_∞ norms, we develop a novel amalgamated ranking metric that evaluates each treatment’s overall and worst performance across all outcomes. Here, weights represent the importance of each outcome to the decision-maker, obtained either by expert opinion or patient preferences. We further extend our method by establishing concrete mathematical criteria to determine whether a unique treatment or a set of treatments should be recommended as the ‘best’ compromise solution across outcomes.

Results We illustrate the use of our approach through a network comparing seven treatment classes for chronic plaque psoriasis in terms of their efficacy and safety. In this example, the outcome-specific treatment hierarchies are notably conflicting, as the most efficacious treatments demonstrate reduced performance in terms of safety. By applying the proposed method, we resolved these conflicts and obtained both an amalgamated treatment hierarchy and the best compromise solution that offers optimal balance between efficacy and safety.

Conclusions The proposed framework provides a novel method for generating amalgamated treatment hierarchies and resolving conflicting hierarchies across outcomes. It identifies either a unique solution or a set of treatments, offering clear guidance on the best compromise. By incorporating outcome weights, our method also supports decision-making based on expert opinion or patient preferences.

4: A Novel Approach for Modelling Components of Complex Interventions in Network Meta-Analysis

Tianqi Yu¹, Anna Chaimani²

¹Center of Research in Epidemiology and Statistics, Université Paris Cité, France

²Oslo Center for Biostatistics and Epidemiology, Department of Biostatistics, University of Oslo, Oslo, Norway

Background Complex interventions consisting of multiple potentially interacting components are increasingly encountered in many health domains. Synthesis of data on complex interventions within networks of studies through component network meta-analysis (CNMA) is advantageous since the different components are shared across several interventions and this allows estimating the effects of the individual components. The majority of NMAs involving complex interventions assume that the effects of the individual components are additive; namely for an intervention consisting of components A and B they assume that $\text{effect}(A+B)=\text{effect}(A)+\text{effect}(B)$. This approach often lacks clinical relevance since it ignores the potential interactions leading to synergistic or antagonistic effects between the components. On the other hand, models that incorporate interactions are more flexible but face challenges with limited statistical power.

Methods We introduce a novel CNMA approach that uses ideas from mediation analysis and models complex interventions by specifying mediating pathways that capture the cumulative and sequential effects of the different components on the outcome. The primary assumption is that in studies combining multiple components, there exists a pathway of effects through which one component influences the outcome directly and/or indirectly via other components. For example, for an intervention comprising components A and B, B has a direct effect on the outcome, while A impacts the outcome both directly and indirectly through B. In interventions with three or more components, a hierarchical framework ranks components by their relative “strength”. “Stronger” components affect “weaker” ones, which mediate their effects on the outcome. These relationships are mathematically expressed using recursive equations that decompose the total effect into direct and mediated effects. The method can be implemented using both frequentist and Bayesian frameworks. In the frequentist setting, iterative algorithms such as Newton-Raphson or BFGS are used for estimation, while in the Bayesian framework, estimation is achieved through MCMC methods.

Results We illustrate our method using data from 56 randomized controlled trials of psychological interventions for coronary heart disease. Compared to standard and additive models, the proposed approach yielded more precise estimates while better capturing the interactions between components.

Conclusion Our approach offers a robust and flexible framework for modeling mechanisms in complex interventions, addressing key limitations of existing methods. However, the identification of plausible pathways requires collaboration with domain experts to ensure clinical relevance.

5: The Impact of using Different Random Effects Models in Meta-Analysis

Kanella Panagiotopoulou¹, Soodabeh Behboodi², Jennifer Zeitlin², Anna Chaimani³

¹Université Paris Cité, Center of Research in Epidemiology and Statistics, Inserm, Paris, France

²Université Paris Cité, Inserm, National Research Institute for Agriculture, Food and the Environment, Centre for Research in Epidemiology and Statistics, Obstetrical Perinatal and Pediatric Epidemiology Research Team, Paris, France

³Oslo Center for Biostatistics and Epidemiology, Department of Biostatistics, University of Oslo, Oslo, Norway

Background In the majority of published meta-analyses, the random effects model is used to allow for heterogeneity across studies assuming a normal distribution for the study effects. It has been claimed previously, that under certain conditions, the between-study normality assumption might be suboptimal. Previous simulation studies that compared several alternative random effects meta-analysis models with the normal model found small differences in bias but larger differences in coverage probability and precision. To date, the real impact on the meta-analytic results and the potential benefit from using alternative random-effects models remain unclear.

Methods To investigate the impact of adding flexibility to the between-study distribution, we used a meta-analysis of 65 cohort studies comparing the cognitive functioning between preterm- and term-born children. Despite the presence of heterogeneity among studies, none of the traditional approaches, such as subgroup analysis and meta-regression, succeeded in identifying the factors that may cause it. We compared the results between 18 different random effects models: a) models based on skewed extensions of the normal and the t-distribution, b) models based on mixtures of distributions, and c) models based on Dirichlet process (DP) priors. We also evaluated the potential of non-normal models to give insight in the true distribution of the underlying effects. Sensitivity analyses on prior distributions and on key model parameters were also conducted.

Results We found small differences in the estimation of the mean treatment effect but larger differences for the between-study variance. Skewed and t-distribution models gave a negatively skewed, heavy-tailed and highly peaked posterior distribution for the random effects. This was in line with the results from models incorporating a test for outliers which suggested the presence of two outlying studies. Models using DP priors revealed the presence of two main clusters of studies suggesting that probably the most important effect modifiers are the level of birth prematurity, the use of matched or unmatched data and the type of Intelligence Quotient (IQ) assessment test. The potential impact of these characteristics together had not been considered in the original meta-analysis. Other mixture models, such

as mixtures of two normal or t-distributions, appeared less informative.

Conclusion Our study highlights that using various random effects models might not affect materially the summary estimates but may assist to explain the observed heterogeneity and provide better insights into the distribution of the underlying effects and the interpretation of the findings.

Prediction / Prognostic Modelling 2

Monday, 2025-08-25 14:00 - 15:30, ETH E23

Chair: Ewout Steyerberg

1: Assessing Time-Dependent Discrimination of Prognostic Model with Intercurrent Treatment: Use of Multi-State Modelling for Inverse Probability Censoring Weighting

Loïc Vasseur^{1,2}, Derek Hazard³, Nicolas Boissel², Martin Wolkewitz³, Jérôme Lambert^{1,4}

¹Epidemiology and Clinical Statistics for Trials & Real-world evidence Research (ECSTRRA), Saint Louis Research Institute, UMR1342, INSERM, Université de Paris, France

²Adolescent and Young Adult Hematology Unit, Saint Louis University Hospital, Assistance Publique-Hôpitaux de Paris (AP-HP), Paris, France

³Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany

⁴Biostatistics and Medical Information Department, Saint Louis University Hospital, AP-HP, Paris, France

Introduction Prognostic markers may be used to inform clinical decision about treatment. As a consequence, it is difficult to detangle the prognostic value of the marker from the treatment effect especially when treatment is started after baseline. Following the estimand framework, a solution is to evaluate discrimination using an “hypothetical” strategy where treatment would never have been started, by censoring follow-up at the time of intercurrent treatment.

When assessing discrimination performances for time to event outcome at a given time t , time-dependent receiver area under the receiver operating characteristic curve is frequently used, with inverse probability censoring weighting (IPCW) to account for censoring before t . Existing methods for IPCW can be marker-independent or marker-dependent but always assume a unique censoring mechanism.

Our aim was to consider the differential impact of the marker on censoring due to loss of follow-up and censoring due to intercurrent treatment when evaluating discrimination performances of a baseline prognostic marker which could guide treatment decision.

Methods We used the Inverse Probability of Censoring Weighting (IPCW) estimation of Cumulative/Dynamic time-dependent AUC to assess discrimination.

Through a time-inhomogeneous multi-state model where the two censoring processes (due to loss of follow-up or intercurrent treatment) correspond to two distinct states, censoring is allowed to depend differentially on the marker (IPCW^{MSM}). This method is compared to IPCW calculations where censoring due to intercurrent treatment or loss to follow-up are combined and considered either marker-independent using a Kaplan Meier estimate of censoring distribution (IPCW^{KM}) or marker-dependent using a Cox proportional hazard model (IPCW^{Cox}). Through simulations we compared bias, coverage probability, and Root Mean Square Deviation (RMSD) of the 3 methods.

Finally, the different methods were illustrated in a cohort of patients treated for acute leukemia to assess the prognostic value of European Leukemia Net 2022 classification, a baseline risk stratification marker, with allogenic hematopoietic cell transplantation in first complete remission as the intercurrent treatment.

Results In the simulation study, time-dependent AUC estimated with IPCW^{MSM} showed reduced bias, better coverage, and lower RMSD, compared to IPCW^{KM} and IPCW^{Cox} .

In the motivating example, the methods provided differing results, which could lead to inconsistent conclusions.

Conclusion When a prognostic marker is used to guide treatment, if we aim to produce unbiased estimate of discrimination performances in the absence of treatment, separating censoring mechanisms due to loss of follow-up and intercurrent treatment using multi-state modelling to estimate IPCW in the calculation of time-dependent AUC is recommended.

2: Evaluating the Discrimination of Prediction Models for Recurrent Medical Events

Thomas Joel Spain¹, Alexandra Hunt¹, Hein Putter², Victoria Watson³, Laura Bonnett¹

¹University of Liverpool, United Kingdom

²University of Leiden, Netherlands

³Phastar, United Kingdom

Background

Clinical prediction models combine multiple pieces of patient information to predict a clinical outcome for individuals with underlying health conditions. Evaluating a model's ability to distinguish between those who experience the outcome and those who do not, referred to as

discrimination, is a key step in model development.

Prediction models are often developed using logistic regression, time-to-event methods, or increasingly, machine learning. Tools and methods are available to assess discrimination and calibration in these models. However, many medical conditions, such as recurrent seizures in epilepsy or repeated asthma exacerbations, are characterized by repeated episodes of the same type. While methodology and tools exist to evaluate the fit and calibration of recurrent event prediction models, there are currently no approaches to evaluate discrimination for these models.

Methods

We propose an alternative concordance statistic, which evaluates predicted and observed event counts, and demonstrate it using simulated data that reflects annual, monthly, and weekly repeated medical episodes. We present R code embedding C++ to minimize computation time, which includes methodology to calculate confidence intervals for the concordance statistic using the jackknife resampling method. This flexibility allows users to tailor the analysis to their specific modelling needs. The code will ultimately be incorporated into an R package to enhance accessibility for researchers.

Results

Embedding C++ within the R code significantly improved computational efficiency. The original R code, when evaluating discrimination for a prediction model developed as part of the PRISE study, required over 24 hours to run. In contrast, the optimized R code embedding C++ completed the same evaluation in under a second. Detailed results on the simulated data, including comparisons across annual, monthly, and weekly event frequencies, will be presented.

Conclusions

This work addresses a critical gap in evaluating discrimination for recurrent event prediction models. The proposed methodology and C++-embedded code provide researchers with a practical and efficient tool. This should ensure that prediction models for recurrent medical episodes can be developed and validated to the same standards as those required by the TRIPOD reporting guidelines, and thus meeting best statistical practice.

3: A Novel Dynamic Prediction Model Based on Interpretable Deep Learning Using Restricted Mean Survival Time

Pansheng Xue, Zheng Chen

Southern Medical University, China, People's Republic of

I. Introduction

In the field of healthcare, survival prediction models have exhibited immense value in clinical practice. Most existing models utilize survival probability as the prediction outcome, making it challenging to answer an important question: “How much longer will the patient live or live without experiencing disease progression?” Thus, a promising clinical time-to-event measure—the restricted mean survival time (RMST) is recommended. In dynamic prediction, RMST can be extended to conditional RMST (cRMST) to adjust the life expectancy on the basis of the time a patient has already survived. Especially in chronic progressive diseases such as Alzheimer’s disease, the patient’s life expectancy can be reflected and updated, which is more intuitive to interpret.

II. Methods

Neural networks have shown powerful capabilities in prediction task but limited by their inherent non-interpretability. Therefore, we developed a novel interpretable RMST dynamic prediction model called *Dynamic-DeepRMST*, in which cRMST is used as the prediction outcome and integrates deep learning techniques. To enhance both accuracy and interpretability, we modified the Transformer model to effectively capture longitudinal data representations for prediction and to transparently quantify the input output relationship.

III. Results

Compared with existing RMST static and dynamic regressions, the proposed model demonstrated superior performance on concordance index and mean absolute error under different simulation scenarios. Furthermore, the model was applied to Alzheimer’s Disease Neuroimaging Initiative (ADNI) data to explain the impact of covariates on survival time from both individual-level and population-level analyses, providing insights for clinical prognosis. Our model not only reveals how longitudinal changes in covariates differentially affect survival time but also identifies the relevance among covariates and their importance in survival prediction evolving over time.

IV. Conclusion

Considering the nature of chronic progressive diseases such as regular follow-ups to track disease progression, we developed a dynamic prediction model based on a more intuitive time-scale-based measure, i.e. RMST, than survival probability. By modifying the Transformer architecture, our model can effectively capture the longitudinal covariate trajectory and inherent relevance while avoiding parametric structural constraints, and achieves comprehensive interpretability by quantifying the input output relationship, providing prognosis insights at both the individual and population levels. To our knowledge, this paper is the first in which an interpretable deep learning dynamic prediction model was investigated.

4: Correcting for Differential Diagnosis Bias Across Protected Attributes in Clinical Prediction Models using Cancer Stage Information and Causal Inference

Jose Benitez-Auriolles¹, Ricardo Silva², Matthew Sperrin¹

¹University of Manchester, United Kingdom

²University College London, United Kingdom

Background Recently, more clinical prediction models are developed using large datasets from routine clinical practice, such as electronic health records. These usually have larger sample sizes and are more representative of the general population, but do not have the same data quality assurances as 'traditional' clinical studies. Particularly, underdiagnosis in routine care is a concern. In England, around 30% of people with type 2 diabetes or hypertension are undiagnosed, and models trained on data from clinical practice will underestimate the overall incidence of these conditions. Differential underdiagnosis happens when a patient's characteristics affect their likelihood of diagnosis. If these characteristics are protected attributes like gender, ethnicity, or socio-economic status, clinical prediction models can exacerbate inequalities by diverting resources away from underserved groups to those already better serviced. Differential underdiagnosis is hard to address, as it is not easily measured. We propose a novel method to correct for differential underdiagnosis in cancer prediction models.

Methods In epidemiology, underdiagnosis in cancer is often indirectly measured through diagnostic delay, as some underserved groups are sicker at the time of diagnosis. If there are quantitative markers of disease progression, these could be used in order to understand which groups are diagnosed later, and correct for this. We show that this is possible in the specific case of cancer stage, assuming that all people with late-stage cancer have the same probability of being tested and diagnosed. We take a causal longitudinal approach, defining our estimand as the counterfactual patient-level risk of being diagnosed in a world in which the probability of being tested is not affected by baseline patient characteristics. By leveraging the difference in the ratio of early to late-stage cancer diagnoses across groups, we can estimate the relative probability of an individual with early-stage cancer to get tested compared to a reference population. This estimate can be used, in turn, to appropriately adjust the prediction scores of groups commonly diagnosed at later stages.

Results We provide theoretical proofs of the identifiability of these counterfactual predictions, and show how to estimate them in practice. We will use a simulation to evaluate the method and benchmark it against alternative approaches.

Conclusion This work has potential applications in cancer screening, particularly in considerations of fairness in early detection. Further work will explore alternative applications and extend the concept to continuous, instead of binary, markers of disease progression.

Joint / Longitudinal Modelling

Monday, 2025-08-25 14:00 - 15:30, ETH E21

Chair: Michael Crowther

1: Assessing Surrogacy from Joint Modelling and Mediation Analysis when Surrogates are Either Censored Event Times or Longitudinal Biomarker: Cancer Application

Virginie Rondeau¹, Quentin Lecoent², Catherine Legrand³

¹Department of Biostatistics, Bordeaux Population Health Research Center, INSERM U1219, Université de Bordeaux, Bordeaux, France

²Department of Public Health Sciences, University of Chicago, Chicago, IL, USA

³ISBA/LIDAM, UCLouvain, Louvain-la-Neuve, Belgium

Background Before a candidate surrogate endpoint can be used in a clinical trial, it must be statistically validated. Joint modelling is a powerful approach to better understand the treatment effect when time-to-event and longitudinal endpoints commonly co-occur.

The aim of our approach is to decompose the total treatment effect on the final endpoint into a direct treatment effect and an indirect treatment effect mediated through a carefully constructed mediation path with a longitudinal mediator or a time to event mediator.

Methods We focus on the cases where the final endpoint is a time-to event endpoint (such as time-to-death) and the surrogate is either a time-to-event or a longitudinal biomarker. Two new joint models were proposed depending on the nature of the surrogate. A mediation analysis is proposed to decompose the total effect of the treatment on the final endpoint as a direct effect and an indirect effect through the surrogate. The ratio of the indirect effect over the total effect of the treatment on the final endpoint can be computed from the parameters of the model and used as a measure of surrogacy. Inference is based on maximization of the penalized partial likelihood approach in the framework of the proposed joint models and has been evaluated via a large scale simulation study.

Results

We present the application to survival mediation analysis using real datasets in oncology with or without a meta-analytic nature of the data in order to quantify the proportion of treatment effect through the surrogate. The proposed mediation analyses study the time-to-relapse as

a surrogate of overall survival in gastric cancer and the tumor size as a surrogate biomarker of overall survival in colorectal cancer.

The R-package `frailtypack` available on <https://CRAN.R-project.org/package=frailtypack> provides a user-friendly implementation of the above estimation and inference procedure.

Conclusion We proposed a valuable tool to inform decision-making and advance our understanding of different treatment effects in mediation analysis with either a longitudinal mediator or a time-to-event mediator for a final survival endpoint.

- Q. Lecoent, C. Legrand, and V. Rondeau. Time-to-event surrogate endpoint validation using mediation and meta-analytic data. *Biostatistics*, 2022.
- Q. Lecoent, C. Legrand, and V. Rondeau. Validation of longitudinal marker as a surrogate using mediation analysis and joint modeling: evolution of PSA as a surrogate for DFS. *Biometrical Journal*, 2024.
- Q. Lecoent, C. Legrand, and V. Rondeau. Tutorial for Surrogate Endpoint Validation Using Joint Modeling and Mediation Analysis. <https://doi.org/10.48550/arXiv.2502.08443>

2: Joint Modelling Using Semiparametric Accelerated Failure Time Approaches: Application to Health-Related Quality of Life Analysis

Ding Ma¹, Patrick Maher¹, Andrew Martin^{1,2}

¹ULTRA Team, Centre for Clinical Research, The University of Queensland, Australia

²Australasian Gastro-Intestinal Trials Group (AGITG)

Background Joint models (JMs) are well established for incorporating longitudinal biomarkers in time-to-event outcome prediction, but may be overlooked for predicting longitudinal health-related quality of life (HRQoL) outcomes in favour of linear mixed models (LMMs). LMMs rest on a missing-at-random assumption that may be uncertain in oncology trials where HRQoL assessments cease at progression. JMs may provide an advantage here, however popular software restricts the time-to-event component to proportional hazards (PH) models or simple parametric models, limiting their applicability to trials like AGITG INTEGRATE IIa (I2a). We have previously investigated a JM (R package `JSM`) with a semiparametric proportional odds (sPO) sub-model which yielded promising results. We now propose two flexible JM approaches that incorporate semiparametric accelerated failure time (sAFT) sub-models.

Methods The first approach¹ used Bernstein polynomials to approximate the baseline hazard function, employing rescaling strategies to enhance computational stability. The second approach² adopted Gaussian basis functions to approximate the baseline hazard, and incorporated a roughness penalty with an automatic smoothing parameter selection procedure. We fitted the two JM models via a Bayesian method using uninformative priors for most parameters, except for those involved in the roughness penalty in the second approach. We applied both approaches to I2a.

Results We obtained stable estimates for I2a using the first approach that aligned with the original I2a results. Estimates from the second approach exhibited instability, indicating areas that require further refinement.

Conclusion The two approaches of sAFT sub-models represent novel alternatives to the proportional hazards models or simple parametric models for use in JMs with HRQoL, however challenges remain fitting the second approach. The dynamic range of the accelerated failure time term required intensive computation via posterior draws. Along with this, the MCMC computation involved many parameters with dispersed initial estimates (due to uninformative priors) thereby reducing stability. We are working to address the instabilities and will present our ongoing efforts.

References

- [1] Panaro RV. spsurv: An R package for semi-parametric survival analysis. arXiv preprint arXiv:2003.10548.2020.
- [2] Ma D, Ma J, Graham PL. On semiparametric accelerated failure time models with time-varying covariates: A maximum penalised likelihood estimation. Stat Med. 2023;42(30):5577–5595.

3: Using Joint Models to Assess Delayed Initiation of Salvage Therapy Following Biochemical Recurrence for Prostate Cancer.

Jeremy M G Taylor¹, Dimitris Rizopoulos², Lukas Owens³

¹University of Michigan, United States of America

²Erasmus Medical Center, The Netherlands

³Fred Hutchinson Cancer Center, United States of America

Prostate cancer patients who undergo prostatectomy are closely monitored for recurrence and metastasis using routine prostate-specific antigen (PSA) measurements. When PSA levels rise, this is called Biochemical Recurrence and at that point salvage therapies (ST) are considered to decrease the risk of metastasis. However, if the likelihood of metastases in

the near future is thought to be low and due to the side effects of these therapies, patients may wish to delay the initiation of salvage therapy. A possible dynamic treatment strategy is that patients delay starting ST until their PSA values rises above a higher threshold. In this work, we use data from Memorial Sloan Kettering Cancer Center to estimate the risk of metastasis under such strategies. Due to the observational nature of this data, we face the challenge that PSA is simultaneously a time-varying confounder and an intermediate variable for salvage therapy. We specify a joint longitudinal survival model for the PSA trajectory, the hazard of metastases and the hazard of death from other causes. The model also incorporates a counterfactual framework. The model is estimated and the predictions for individual patients are made using a Bayesian approach, implemented via the R package JMbayes2.

4: Faster Estimation of Quasi-Monte Carlo Methods for Joint Models of Multivariate Longitudinal Data and Penalized Cox Regression

Adeboye Azeez, Colin Noel

University of Free State, South Africa

The estimation of joint models for time-to-event and multivariate longitudinal data presents significant computational challenges, particularly when utilizing Monte Carlo simulations. This study compares classical Vanilla Monte Carlo (MC) methods with two Quasi-Monte Carlo (QMC) techniques, Sobol and Halton sequences, in the context of joint models incorporating penalized Cox regression. The QMC integration framework extends the Monte Carlo Expectation Maximisation approaches that are commonly adopted. The motivation behind QMC sequences to improve convergence speed and computational efficiency by distributing nodes more uniformly. Simulations and a clinical dataset application demonstrate that QMC methods outperform Vanilla MC in terms of convergence speed, computational efficiency, and accuracy for all sample sizes, offer a distinct convergence speed advantage. By reducing variance and accelerating convergence, QMC methods provide a more efficient alternative for fitting complex joint models with penalized regression, especially in high-dimensional settings. The findings highlight the advantages of QMC methods in improving the practical application and computational feasibility of joint modelling approaches.

5: Regularization and Flexible Methods to Improve Complete Cancer Prevalence Predictions in the Prevalence Incidence Analysis Model

Fabrizio Di Mari¹, Roberta De Angelis², Therese ML Andersson³, Enoch Yi-Tung Chen³, Silvia Rossi², Paul W Dickman³, Roberto Rocci¹, Mark Clements³

¹Sapienza University of Rome, Italy

²Italian National Institute of Health, Italy

³Karolinska Institutet, Sweden

Cancer prevalence is one of the primary measures used to assess the impact of the disease on a population. It helps the healthcare system quantify the burden of cancer and allocate resources to improve patient care. Prevalence is defined as the proportion of individuals with a current or past diagnosis of cancer within a population at a specific time. The Prevalence Incidence Analysis Model (PIAMOD) is commonly used to estimate the prevalence of an irreversible disease, such as cancer, in population-based studies. It primarily relies on modeling two key estimands, incidence and net survival, which are combined in a back-calculation method to estimate prevalence. Incidence rate is specified as an Age-Period-Cohort (APC) model using Poisson regression, while net survival is estimated within a relative survival framework. However, the APC model selection is based on a stepwise procedure that does not consider any measures of generalization performance. Additionally, net survival is estimated either non-parametrically, which results in high variability at the tails, or using a Weibull mixture cure model, which has been shown to lack sufficient flexibility to fit the observed data. These issues are significant concerns when the primary objective is to assess the future burden of a disease.

To address these drawbacks, we propose a Least Absolute Shrinkage and Selection Operator (LASSO)-driven model selection procedure for selecting the APC model. The strength of the penalization is chosen to minimize prediction error in the last observed years, left out during the training phase. For estimating net survival, we use flexible parametric survival cure models, which have been shown to adapt much better to the data than the Weibull mixture cure model and have less variability at the tails than the non-parametric estimator. We compared the novel and classical procedures using data from colon cancer patients diagnosed between 1958 and 2019 in Sweden. Furthermore, we compared the predicted incidence cases and cause-specific mortality cases with observed data up to 2023 using public sources. Our proposed method performed better than the classical approach across most of the validation measures considered.

Key References

Verdecchia, A., De Angelis, G., Capocaccia, R. (2002). *Estimation and projections of cancer prevalence from cancer registry data*. Statistics in Medicine, 21(22), 3511–3526.

Abstracts of Contributed Talks

Andersson, T. M. L., Dickman, P. W., Eloranta, S., & Lambert, P. C. (2011). *Estimating and modelling cure in population-based cancer studies within the framework of flexible parametric survival models*. BMC Medical Research Methodology, 11(1), 96.

Estimands: Causal and Multiple Imputation Approaches

Monday, 2025-08-25 16:00 - 17:30, Biozentrum U1.131

Chair: Ian R White

1: The Role of Post Intercurrent Event Data in the Estimation of Hypothetical Estimands in Clinical Trials

Jonathan Bartlett

London School of Hygiene & Tropical Medicine, United Kingdom

Background Estimation of so-called hypothetical estimands in clinical trials has historically not made use of data that may be collected after the intercurrent event (ICE). Some recent papers (e.g. Lasch et al 2023) have shown that such data can be used for estimation of hypothetical estimands using causal inference estimators (e.g. G-formula and G-estimation), and that these can be more precise (and thus have higher power) compared to using estimators that only use data before the ICE. This raises the question of whether trials should routinely be using such post-ICE data when estimating hypothetical estimands.

Methods We critically examine missing data and causal inference methods for estimating hypothetical estimands which do, and do not, make use of post-ICE data, and in particular what assumptions must be made in order for precision to be increased. We derive asymptotic variance expressions for the estimators to quantify the potential gain in precision and power. Simulations are used to verify these.

Results We first show that G-formula and G-estimators for hypothetical estimands are identical in some important special cases, and as such in these cases their precision/bias properties are identical. We show that estimators that use post-ICE data can only increase precision by assuming the effect of the ICE on outcome is not modified by those variables that affect the occurrence of the ICE and the final outcome, and we argue that such an assumption will often not be plausible. Moreover, our asymptotic variance calculations reveal that in practice the gain in power to detect an effect, even if such assumptions are made, will typically be modest.

Conclusion Given that the assumptions required to improve precision will, in our view, often not be plausible, and that even if one makes such assumptions, the gain in precision will typically be modest, we recommend that in general estimation of hypothetical estimands

continue to be performed using estimators which only use data up until the occurrence of the ICE.

Reference Lasch F, Guizzaro L, Pétavy F, Gallo C (2023). A Simulation Study on the Estimation of the Effect in the Hypothetical Scenario of No Use of Symptomatic Treatment in Trials for Disease-Modifying Agents for Alzheimer's Disease, *Statistics in Biopharmaceutical Research*, 15:2, 386-399, DOI: 10.1080/19466315.2022.2055633

2: Bringing Together Estimands, Causal Inference and Pharmacometric Modeling and Simulation

Christian Bartels¹, Manuela Zimmermann¹, Siyan Xu², Neva Coello¹

¹Novartis Pharma AG, Basel, Switzerland

²Novartis Institutes for BioMedical Research, Cambridge, Massachusetts, USA

The estimand framework described in the ICH E9 (R1) addendum, causal inference, and pharmacometrics modeling and simulation (M&S) are tools that can help to frame and answer causal questions about the efficacy and safety of clinical products. We aim at integrating these approaches to leverage their respective advantages. The estimands framework translates clinical questions into precisely defined quantitative measures, the estimands. The framework considers post-randomization intercurrent events (IE) that may confound the relationship between the drug intake and the response. Causal inference offers theory, assumptions and methods to express estimands in terms of statistical quantities that can be inferred from observed data. However, often the available data on the clinical endpoint of interest is limited such that an estimand might not be identifiable. This is particularly true when the question of interest concerns a hypothetical estimand, interpolation, e.g. an intermediate dose level, or extrapolation, e.g. to a pediatric population. Pharmacometrics modeling and simulation aims at bridging this gap by taking advantage of possibly available longitudinal pharmacokinetic (PK) and pharmacodynamic (PD) data and external knowledge on the drug and disease in the form of semi-mechanistic models.

We have shown that standard nonlinear mixed effects modeling (NLME M&S) is an implementation of a well-known method in causal inference, standardization. Standardization corrects for confounding by analyzing and combining results from groups of similar patients. In the NLME M&S implementation, conditioning occurs on individual parameters described by the random effects of the NLME model.

Currently we are evaluating the potential and limitations of NLME M&S to correct for

confounding. The focus is on studies in which intercurrent events related to either efficacy or tolerability may lead to deviations from the assigned treatment schedule, and on hypothetical estimands of the efficacy at the end of the trial assuming adherence to a given treatment regimen.

We found that if the PK and PD of a drug are well understood and the clinical studies provide rich PK and PD data, NLME M&S is reliable, even with unobserved confounders affecting both the outcome and the dosing. For the more common situation that PK is well understood and supported by data, but only limited data or knowledge is available for the PD, we show that NLME M&S may provide reliable estimates in some situations but fails in others.

3: Nonparanormal Adjusted Marginal Inference

Susanne Dandl, Torsten Hothorn

Epidemiology, Biostatistics & Prevention Institute, Universität Zürich, Switzerland

Introduction Treatment effects are typically defined as measures comparing the marginal outcome distributions observed in two or more study arms to assess the efficacy of a novel therapy. Unbiased marginal estimates can be obtained under proper randomisation without adjustments, but previous work showed that covariate adjustments can improve precision even in randomised controlled trials. For non-collapsible effect measures - such as Cohen's d in linear regression, log-odds ratios in binary logistic regression or log-hazard ratios in the Cox model - conditioning on covariates changes the interpretation of treatment effects, making conditional and marginal effects incomparable.

Method We propose a novel nonparanormal model formulation for adjusted marginal inference allowing the estimation of the joint distribution of outcome and covariates featuring the intended marginally defined treatment effect parameter whose interpretation is unaffected by adjustment. Corresponding marginal distributions are modelled by transformation models allowing broad applicability to diverse outcome types (including, non-normal continuous, binary, ordinal or right-censored survival outcomes). For the special case of Cohen's d under normally distributed outcomes and covariates, we present a closed-form expression of the standard error under covariate adjustment to investigate the potential for sample size reductions theoretically.

Experiments: We evaluated the ability of our proposed method to increase precision under different prognostic strengths of covariates and under different numbers of noise variables in

a simulation study. We compared the results with a model conducting unadjusted marginal inference, conditional inference, and previously proposed adjustment methods for marginal inference.

Results Our proposed approach obtained unbiased parameter estimates of marginally defined parameters with covariate adjustment leading to reduced standard errors, and, thus, narrower Wald confidence intervals, compared to unadjusted marginal inference. The advantages became more pronounced as the prognostic strength increased. Adding noise variables had no large effects on the adjusted parameter estimates.

Conclusion Overall, this reveals the potential of our method to bypass the problems induced by non-collapsibility of practically important treatment effect measures such as Cohen's d, log-odds ratios, or log-hazard ratios. The approach allows for extensions to estimate heterogeneous treatment effect estimates potentially under observational data offering exciting opportunities for further research.

4: Estimation of Effects with Treatment Policy Handling for Binary Outcomes using Multiple Imputation

Sunita Rehal

GSK

Introduction ICH E9 (R1) makes a distinction between what is to be estimated (the estimand) and how to estimate it (the analysis). The estimand should include detail about post-baseline events (intercurrent events) and how they will be handled. The impact intercurrent events (IEs) have on the effect targeted will depend on the handling strategy chosen. A common handling strategy is treatment policy. This means patients' outcome measures post-IE are deemed clinically relevant to estimate the treatment effect and information after the IE should continue to be collected and included in the final analysis. In this setting, the occurrence of missing data post-IE complicates estimation. Multiple imputation is one approach to account for post-IE missing information and research has been done so far for continuous, time-to-event and recurrent event endpoints. We show how this type of estimand can be estimated in the binary setting while accounting for pre- and post-IE information and in the presence of missing data.

Methods We present the results from a simulation study where we create binary repeated outcomes investigating four different models: a basic missing at random (MAR) model that does not account for the occurrence of the IE, a pre- and post-IE model, a pattern linear

model and pattern full model.

Results Using a basic MAR model showed it was the most biased model which increases as the rate of the IE increases and as the missing data increases. Models that attempt to account for the IE appear to work well, provided there is enough post-IE information recovered, but runs into computational issues for the most complex model.

Discussion In general, basic MAR models are poor options for estimating effects on binary outcomes that use a treatment policy strategy for IEs. The choice of the most appropriate model will depend on the disease area and the expected rate of collecting post-IE information and it is critical to consider these when choosing an estimation method.

5: Can Treatment Effect Testing in Trials with Intercurrent Events be Nearly Assumption-Free?

Georgi Baklicharov, Kelly Van Lancker, Stijn Vansteelandt

Ghent University, Belgium

Intercurrent events, such as treatment switching, rescue treatment, and truncation by death, pose significant challenges to the interpretation of treatment effects in randomized clinical trials. Intention-to-treat analyses often fail under these conditions, potentially resulting in misleading conclusions. Existing methods, including hypothetical estimands and survivor average causal effects, address some challenges but rely on strong assumptions, are prone to positivity violations, and struggle with time-varying confounders. In this talk, I will present a novel methodology for analyzing longitudinal clinical trial data impacted by intercurrent events. Our approach does not require data on time-varying confounders and does not exclude positivity violations on the intercurrent events. It relies on a weak structural assumption about the occurrence of intercurrent events and is found to deliver only small bias under its violation. We propose asymptotically efficient, model-free tests of the null hypothesis of no treatment effect, which make use of data-adaptive nuisance parameter estimates. In the context of randomized experiments, we moreover propose asymptotically efficient tests in a subclass of tests that have greater robustness properties. The methodology's empirical performance is demonstrated through simulation studies and the re-analysis of a recent diabetes trial, which is complicated by truncation due to death.

Clinical Trials and Regulatory Issues

Monday, 2025-08-25 16:00 - 17:30, Biozentrum U1.141

Chair: Marcia Rueckbeil

1: Compatible Effect Estimation and Hypothesis Testing in Drug Regulation

Samuel Pawel, Leonhard Held

University of Zurich, Switzerland

The two-trials rule in drug regulation requires independent statistically significant results from two pivotal trials to demonstrate drug efficacy. However, it is unclear how effect estimates from two trials should be combined to quantify the drug effect. Combination with meta-analysis can lead to situations where the two-trials rule is not satisfied, but the meta-analytic confidence interval excludes the value of no effect. Here we show how the two-trials rule and meta-analysis can be cast in the framework of combined p-value functions, where they are variants of Wilkinson's and Stouffer's combination methods, respectively. We show how compatible combined p-values, effect estimates, and confidence intervals can be obtained, and derive them in closed-form. We also investigate Edgington's, Fisher's, Pearson's, and Tippett's p-value combination methods. We find that when both trials have the same true underlying effect, all methods can consistently estimate it, although some methods show bias. When the true trial effects differ, Fisher's and Tippett's methods are asymptotically anti-conservative (converging to the more extreme effect), the two-trials rule and Pearson's method are conservative (converging to the less extreme effect), and Edgington's method and meta-analysis are balanced (converging to a weighted average). However, Edgington's method asymptotically always includes the individual effect estimates in its confidence interval while the meta-analytic confidence interval converges to a point. We conclude that all these methods may be appropriate depending on the estimand of interest.

2: Opportunities to Speed Up IVD Adoption and Patient Access in the UK: The Pre-Eclampsia Testing Timeline

Katie Scandrett¹, Joy Allen², Jon Deeks¹, Julia Eades², Ashton Harper², Christopher Hyde³, Yemisi Takwoingi¹, David Wells⁴

¹Department of Applied Health Sciences, University of Birmingham, UK

²Access and Innovation, Roche Diagnostics Limited, UK and Ireland

³University of Exeter Medical School, UK

⁴Institute of Biomedical Science

In-vitro diagnostics (IVDs) are increasingly important in modern healthcare. The use of effective IVDs can substantially improve patient outcomes and there is opportunity for research and design innovation and investment. However, the pathway between the development of a new IVD and widespread adoption in the UK is complex and there are multiple points along the innovation pathway where bottlenecks may occur. Using the Roche Elecsys sFlt-1/PIGF ratio test to diagnose pre-eclampsia as an exemplar, we illustrate the full pathway from evidence development to adoption and highlight the challenges and barriers to widespread implementation and patient access in the UK. We will then discuss potential solutions to support more timely access to innovations.

In 2008, a Health Technology Assessment (HTA) outlined the need for accurate biomarkers to diagnose pre-eclampsia. However, it was not until 2016 that there was enough evidence for the National Institute for Health and Care Excellence (NICE) to recommend use of PIGF-based testing to rule out pre-eclampsia. Following the recommendation, there were significant implementation issues in the National Health Service due to funding and procurement barriers. Additional funding was obtained to reimburse the cost of the test, but barriers to implementation persisted until key stakeholders worked together on a national level to prioritise pathway transformation resource. A review of the 2016 NICE guidance began in 2020, and new NICE guidelines published in 2022 continue to endorse use of the PIGF-based testing for both rule-in and rule-out indications.

The evidence generation pathway for an IVD is more complex than for pharmaceuticals and should follow a dynamic, cyclical approach. However, more research is needed to inform the methods for doing so. In particular, further guidance is needed to allow for effective combination of clinical performance and clinical effectiveness data, which takes into account the full value proposition of the diagnostic technology aside from improvements in accuracy. Mandated funding following NICE approval of IVDs should be considered, and key stakeholders should collaborate to identify barriers to adoption, especially given the projected growth of the IVD market in upcoming years.

3: Statistical Review of Regulatory Requirements for AI Diagnosis

Naoki Ishizuka¹, Taro Shibata²

¹Kyoto University Graduate school of Medicine, Japan

²National Cancer Center, Japan

1 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 according to the regulation or the act at first [1]. We will review the latest status so called SaMD: Software as Medical Device in regulation for AI diagnosis with Gastrointestinal endoscopy as an example.

2. 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

3. Result

So far thirteen SaMD have been approved by Pharmaceuticals and Medical Devices Agency in Japan as of February 14, 2025 while five SaMD have been approved by US Food and Drug Agency as of October 1, 2024. There is only one SaMD which was approved in both US and Japan.

According to the package inserts which are available at the site of Japanese Pharmaceutical and Medical Device Agency, 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 510(k) Premarket Notification which are available at the site of US Food and Drug Agency, 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.

4. Conclusion

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. 2020;92(4):801-806.

4: Validating Physiologically-Based Pharmacokinetic Models using the Continuous Ranked Probability Score: Beyond Being Correct on Average

Laurens Sluijterman, Marjolein van Borselen, Rick Greupink, Joanna intHout

Radboud University Medical Center, Netherlands, The

Introduction Physiologically-based pharmacokinetic (PBPK) models are becoming increasingly popular for model-informed drug development (MIDD), prompting the development of several evaluation frameworks and regulatory guidance documents. This guidance is, by design, often of a general and non-specific nature: it is clear what steps should be carried out, but not necessarily how. For example, one of the necessary steps is validation, where the predictions of the model are compared to an external dataset. However, how to carry out this comparison remains unclear and is therefore the topic of this work.

Methods We propose a validation step based on the popular Continuous Ranked Probability Score (CRPS). Contrary to the current standard of only evaluating if the model is on average correct, this metric explicitly measures how well the model captures the *distribution* of the observed data. The CRPS is applicable to both individual-level predictions and virtual population simulations, making it well-suited for PBPK modeling. Additionally, we demonstrate that the average CRPS for the entire dataset can be computed with only a single integral, greatly increasing computational efficiency and thereby facilitating the computation of bootstrap confidence intervals. Lastly, the average CRPS of an individual model can be compared

with the average CRPS of a naïve one-dimensional model to obtain a skill score, an intuitive measure of model performance.

Results We applied the CRPS-based validation approach to compare two PBPK models, showing that it gives much more insightful and robust validation than the current standard. Additionally, we showed that the skill score offers a clear interpretation of the obtained CRPS scores. Surprisingly, both PBPK models produced a negative score.

Conclusion The proposed validation method offers a more quantitative and interpretable approach to PBPK model evaluation. While our focus is on PBPK models, this framework is broadly applicable to other modeling scenarios that involve virtual populations. To facilitate adoption, we provide an accessible online tool that implements the CRPS-based validation method.

5: Conditional Marketing Authorisation Based on the Intermediate Endpoint of a Randomised Clinical Trial: Dual or Co-Primary Endpoints?

Nele Henrike Thomas¹, Xiaofei Liu^{1,2}, Elina Asikanius³, Anika Großhennig¹, Armin Koch¹

¹Hannover Medical School (MHH)

²Federal Institute for Drugs and Medical Devices (BfArM)

³Finnish Medicines Agency (Fimea)

Background/Introduction Conditional marketing authorisation (CMA) is a way of allowing early market access of new drugs addressing an unmet medical need, especially for serious or life-threatening diseases (European Union, 2006). For justifying a CMA, the benefit-risk profile has to be positive and the applicant must be able to provide comprehensive data post-authorisation for converting it to a full marketing authorization (MA). Initial attempts of basing a CMA on promising data from a single-arm trial, complemented by a randomised clinical trial post-authorisation, were notoriously difficult. A recent proposal is to plan a randomised clinical trial with interim analyses and basing the decision for CMA and full MA on an intermediate and the final primary endpoint, respectively. To control the study-wise type I error, the dual primary endpoint concept, which is essentially a Bonferroni-split of the study-wise type I error between the intermediate and final primary endpoint, is applied. We argue that the resulting statistical definition of study success is not in line with clinical and regulatory assessment of the overall trial outcome (Großhennig & Thomas, 2023).

Methods Consistent with European Regulatory Guidance on multiplicity issues in clinical

trials (EMA, 2002), we propose to define the intermediate and final primary endpoint as co-primary – each assessed with a standard group sequential design testing procedure. We illustrate our proposal using an oncological phase III clinical trial with complete remission as intermediate and overall survival as final primary endpoint.

Results/Conclusion:

Based on the example, we demonstrate that our proposal is a valid and flexible alternative that would not increase costs in terms of sample size or timing of the interim analysis for applying a CMA but clearly improves the interpretation of the overall trial outcome.

References

- European Medicines Agency. Committee for proprietary medicinal products. (2002). Points to consider on multiplicity issues in clinical trials. https://www.ema.europa.eu/en/documents/scientific-guideline/points-consider-multiplicity-issues-clinical-trials_en.pdf
- European Union (2006, March 30). Commission regulation (EC) No 507/2006 of 29 March 2006 on the conditional marketing authorisation for medicinal products for human use falling within the scope of Regulation (EC) No 726/2004 of the European Parliament and of the Council. Official Journal of the European Union. <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32006R0507>
- Großhennig, A., Thomas, N. H., Brannath, W., & Koch, A. (2023). How to avoid concerns with the interpretation of two primary endpoints if significant superiority in one is sufficient for formal proof of efficacy. *Pharmaceutical Statistics*, 22(5), 836-845. <https://doi.org/10.1002/pst.2314>

Model Selection and Simulations

Monday, 2025-08-25 16:00 - 17:30, Biozentrum U1.101

Chair: Tim P. Morris

1: A Systematic Review of Variable and Functional Form Selection Methods Used in Covid-19 Prognostic Models

Michael Kammer¹, Marc Y. R. Henrion^{2,3}, Gregor Buch⁴, Georg Heinze¹

¹Medical University of Vienna (Austria)

²Malawi Liverpool Wellcome Research Programme (Malawi)

³Liverpool School of Tropical Medicine (UK)

⁴University Medical Center of the Johannes Gutenberg University Mainz (Germany)

Background The Covid-19 pandemic created a pressing need for accurate predictions of health outcomes related to the disease. Numerous statistical and machine-learning models were developed in response. A study by Wynants et al. (2020) found that nearly all were at high risk of bias. Several members of the STRATOS topic group on variable and function selection (TG2) hypothesized that, in response to this public health emergency, researchers relied on modeling strategies familiar to them, or those they perceived as trustworthy for producing robust results. Consequently, the published models offer a valuable opportunity to examine current practices in variable selection and functional forms in statistical regression models. On behalf of STRATOS TG2, we systematically reviewed the model building approaches used in these papers.

Methods A systematic re-review of published models in the existing database from the study by Wynants et al (2020) was conducted. A detailed protocol comprising inclusion criteria and a structured questionnaire to extract precise information about the modelling strategy used were prespecified and preregistered. The primary focus of our study was on regression-based models and, specifically, on the methods used for variable selection and the incorporation of functional forms. Data extraction based on a full text review was performed independently by two reviewers per paper, followed by consensus-based consolidation.

Results A total of 20 reviewers extracted data from 181 regression-based prognostic models. We observed considerable variability in approaches to variable selection and functional forms, with researchers frequently combining multiple methods. Univariable selection was widely used, often in multi-stage variable selection strategies. Only very few studies accounted for non-linear functional forms or interactions, mostly by splines or multivariable fractional

polynomials. Many papers also reported on statistical inference, e.g. through confidence intervals for model coefficients, but failed to account for additional uncertainty due to model selection. Notably, the existing, if limited, best-practice recommendations for model building were rarely cited. Overall, reporting quality was often poor, making it challenging to precisely determine the modeling strategies applied.

Conclusion Our review demonstrates the reliance of many study authors on simplified modeling strategies or combined modeling strategies that were not properly tested and have unknown statistical properties. These findings underscore the need for clearer, more comprehensive and more accessible guidance on modeling strategies to support practitioners in developing robust, reliable prediction models.

References Wynants L et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ*. 2020

2: Robust Standard Errors for Coefficients of Selected and Unselected Predictors after Variable Selection for Binary Outcomes

Nilufar Akbari¹, Ulrike Grittner¹, Georg Heinze²

¹Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Institute of Biometry and Clinical Epidemiology, Charitéplatz 1, Berlin, 10117, Germany

²Institute of Clinical Biometrics, Center for Medical Data Science, Medical University of Vienna, Spitalgasse 23, Vienna 1090, Austria

This study aimed to investigate if a modified sandwich estimator of the covariance could help achieve valid standard errors for regression coefficients after variable selection. In regression models, predictors can either be pre-specified or selected through data-driven methods. Valid standard errors for regression coefficients can be estimated for pre-specified models based on established theory. However, for data-driven methods, there is no widely accepted way to compute standard errors that encompass the uncertainty of selection. Most commonly the uncertainty in selection is ignored and standard errors for non-selected predictors are treated as zero. The bootstrap may be used to investigate the variability of selections, however, this method is computationally intensive. This highlights the need for a new, simpler approach.

After having obtained promising results from a previous simulation with a continuous outcome, we now investigated regression with binary outcomes. We performed a simulation study in the setting of logistic regression with five true and five noise candidate predic-

tors, using three correlation structures between those predictors and three different marginal event rates. We compared four methods for obtaining standard errors of the regression coefficients of all candidate predictors, regardless of selection. These methods included using the coefficients and model-based variance of the full model, the final selected model with model-based variance and collapsed variances for non-selected predictors, the final selected model combined with the sandwich method modified to supply standard errors for all candidate predictors and the bootstrap variance method.

Results showed that the full model attained standard errors similar to the true sampling standard errors, which were underestimated by those of the selected model with model-based variances, particularly for correlated predictors. The bootstrap methods performed well for non-predictors and strong predictors but slightly worse for weak predictors. The sandwich method's standard errors were close to the bootstrap results for non-predictors and slightly underestimating for true predictors. For correlated predictors the sandwich method performed slightly better than the bootstrap.

In any case, the modified sandwich estimator improved over the common practice of ignoring uncertainty induced by variable selection, but more research is needed to refine the method. In addition to the simulation, we applied the proposed method to a real data example, which produced results consistent with the simulation findings. This work was supported by DFG grants RA-2347/8-1 and BE-2056/22-1 and FWF grant I-4739-B.

3: The “multi-Performance Plot” in Simulation Studies: a Compact Visualisation of Up to Seven Performance Measures Comparing Multiple Statistical Methods

Wang Pok Lo

Centre for Population Health Sciences, Usher Institute, University of Edinburgh, UK

Background Simulation studies can be used to evaluate the performances of statistical methods, such as to determine how accurately or precisely an estimand can be estimated by each method. Commonly used performance measures include bias, empirical standard error (EmpSE), mean squared error (MSE), average model standard error (ModSE), and coverage (Morris *et al.* [2019]). Tables are a natural way to present these measures. However, pattern identification from such tabular presentations may be hindered when their sizes become large. This occurs when (1) many performance measures are assessed; (2) many combinations of simulation parameters are evaluated, such as in full factorial designs, and (3) many statistical methods are evaluated.

Methods A new two-dimensional plot termed a “multi-performance plot” is proposed to simultaneously address all three scenarios. In scenario (1), the plot is generated as follows. For each combination of parameters simulated, the estimated bias and estimated EmpSE are plotted on the horizontal and vertical axes respectively. Points closer to the origin are more desirable. The squared distance from a point to the origin is approximately the estimated MSE. Next, a vertical line, whose length is the estimated ModSE, is drawn downwards from the point. Finally, each point is shaded using an appropriate colour gradient to visualise the estimated coverage. This plot additionally allows the visualisation of two relative performance measures described in Morris *et al.* [2019], namely the relative error in ModSE and relative precision. If scenario (2) and/or scenario (3) arise with scenario (1), the shapes of points, and the colours and types of vertical lines can be varied to reflect different simulation parameters and statistical methods.

Results The utility of the multi-performance plot is illustrated in two examples: the outperformance of a non-parametric method of surrogate endpoint evaluation (Parast *et al.* [2016]), and good performance of a new model accounting for dependent censoring in survival analysis (Deressa and Van Keilegom [2020]).

Conclusion The multi-performance plot displays up to seven performance measures and can complement tabular presentations for easier identification of patterns, especially in large simulation studies.

Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Statistics in medicine. 2019 May 20;38(11):2074-102.

Parast L, McDermott MM, Tian L. Robust estimation of the proportion of treatment effect explained by surrogate marker information. Statistics in medicine. 2016 May 10;35(10):1637-53.

Deressa NW, Van Keilegom I. Flexible parametric model for survival data subject to dependent censoring. Biometrical Journal. 2020 Jan;62(1):136-56.

4: What is the Impact of Increasing Numbers of Auxiliary Variables for an Explanatory Variable in Imputation Models?

Paul Madley-Dowd¹, Rheanna Mainzer², Samantha Ip³, Alexia Sampri³, Carmen Petitjean³, Jonathan Sterne¹, Katherine Lee², Tim Morris⁴, Angela Wood³, Kate Tilling¹

¹University of Bristol, United Kingdom

²Murdoch Children's Research Institute, Australia

³University of Cambridge, United Kingdom

⁴University College London, United Kingdom

Background: When performing multiple imputation to explore an exposure-outcome association, auxiliary variables are included in imputation models, but not the analysis model, to reduce bias and/or improve statistical efficiency. Recent work has provided evidence for an inclusive strategy for including auxiliary variables in the imputation model (i.e., all available) when there is missing data in the outcome variable. Such a finding is not expected to hold for missing data in other variables such as the exposure or confounders. The aim of this work is to explore the impact of increasing numbers of auxiliaries when imputing an exposure or confounder variable under different missingness mechanisms. Given the increasing sample sizes available using electronic healthcare records, and the different sampling strategies employed to ensure computational viability, it is important to explore impacts under different sample sizes.

Methods: We simulated a complete dataset including an exposure, an outcome, a confounder, and between 1 and 100 auxiliary variables that collectively explained between 10% and 60% of the variance in a missing variable. We explored 50% missing data in either the exposure or the confounder variable. We repeated our simulation at sample sizes of 1000, 10,000, and 1,000,000 and explored different missingness mechanisms. We explored scenarios where the exposure, outcome, and confounder were continuous or binary, using multivariable linear regression as the analysis model when the outcome was continuous and multivariable logistic regression when the outcome was binary.

Results: Using a sample size of n=1000 shows that larger numbers of auxiliary variables that weakly predict the incomplete variable can introduce substantial quantities of bias in estimates of exposure-outcome associations. This bias is often reduced when using variables that more strongly predict the incomplete variable, though this depends on the variable type (continuous or binary) for each analysis model variable and the missingness mechanism. Increasing the sample size also reduces the size of the bias.

Conclusions: We caution against using an inclusive strategy of all available auxiliary variables when imputing an incomplete exposure or confounder variable. Our results suggest that careful consideration needs to be given to 1) how predictive an auxiliary variable is of an incomplete exposure or confounder variable that is to be imputed, 2) the variable type (continuous/binary) for each variable in the analysis model, and 3) the assumed missing data mechanism. Our work provides guidance on the importance of these factors at different sample sizes.

Bayesian Methods 1

Monday, 2025-08-25 16:00 - 17:30, ETH E27

Chair: Daniel Sabanés Bové

1: A Bayesian Model for Surrogate Endpoint Evaluation in Mixed Biomarker Patient Populations

Lorna Sophie Kate Wheaton¹, Stephanie Hubbard¹, Sandro Gsteiger², Sylwia Bujkiewicz¹

¹University of Leicester, United Kingdom

²Global Access, F Hoffman-La Roche AG, Basel, Switzerland

Background Surrogate endpoints are increasingly being utilised in clinical trials and for regulatory and reimbursement decision-making, as they can allow for the treatment effects to be measured more quickly than final clinical outcomes. However, surrogate endpoints should be validated before they are used to inform healthcare decision-making, to ensure that the putative surrogate endpoint is truly predictive of the treatment benefit on the final outcome. Traditionally, the strongest level of evidence for a surrogate relationship would be obtained from a meta-analysis of treatment effects on the surrogate endpoint and final outcome utilising all clinical trial data in the relevant clinical setting. However, in several disease areas genetic biomarkers are predictive of treatment effect and thus may also impact surrogacy relationships. Potential differences in the surrogate relationships across biomarker groups can be investigated via subgroup analysis using standard meta-analytic methods for surrogate endpoint validation. However, this could result in insufficient data to make robust conclusions about the strength of the surrogate relationship, as subgroup analyses tend to be under-reported.

Methods We propose an extension to bivariate random-effects meta-analysis (BRMA) to allow for treatment effects to vary across biomarker subgroups by assuming systematic differences in treatment effects (on both outcomes) between biomarker-positive and biomarker-negative patient subgroups. The systematic differences estimated from studies reporting treatment effects in both biomarker subgroups is then used to interpolate treatment effects in the biomarker-positive subgroup from studies only reporting treatment effects in the mixed population. The true treatment effects in the biomarker-positive population from all the trials are then used to estimate the surrogate relationship in the biomarker-positive subgroup.

Results The standard and proposed models when applied to an illustrative example in non-small cell lung cancer (NSCLC) did not provide strong evidence for surrogacy in biomarker-

positive patients. However, the proposed model reduced the width of the credible intervals for the surrogacy parameters by up to 75% (when using limited clinical trial data available early in the drug development) compared to the standard BRMA model applied to subgroups.

Conclusions The developed method can improve precision of the estimates of surrogacy parameters compared to using the BRMA model for subgroups alone. The improvement in precision of the surrogacy parameters was particularly notable at the early drug development stage when data from only few clinical trials were available. We carried out a simulation study, which confirmed the improved precision of the surrogacy parameters achieved via the extended BRMA model.

2: Extending Bayesian Causal Forests for Longitudinal Data Analysis: A Case Study in Multiple Sclerosis

Emma Prevot, Thomas E. Nichols, Chris Holmes, Habib Ganjgahi

University of Oxford, United Kingdom

With the growing availability of large-scale longitudinal studies, such as the UK Biobank and NO.MS dataset [1], there is an increasing need for scalable predictive models that can accommodate long-term outcomes and perform causal inference in longitudinal settings. Bayesian Additive Regression Trees (BART) has gained popularity in causal inference due to its flexibility, scalability, built-in uncertainty estimation, and ability to capture complex, non-linear relationships and interactions without explicit parametric assumptions [2]. However, existing BART-based causal inference models, such as Bayesian Causal Forests (BCF), assume independent outcomes, making them unsuitable for longitudinal settings, where repeated measures within individuals are inherently correlated. Motivated by the NO.MS dataset, which is the largest and most comprehensive dataset on Multiple Sclerosis (MS), comprising more than 34,000 subjects with up to 15 years follow-up, we develop BCFLong, a flexible hierarchical model that preserves BART's strengths while extending it for longitudinal data analysis. Inspired by BCF, we decompose the mean structure into prognostic and treatment effects, and introduce individual-specific random effects, including random intercepts and time-dependent slope. Additionally, to account for heterogeneous variability across individuals, we implement a sparsity-inducing horseshoe prior on the random effects, which adaptively shrinks small coefficients while preserving meaningful signals. This hierarchical structure balances flexibility and regularization, enabling BCFLong to adapt to varying levels of complexity in fixed and random effects.

Simulation results show that BCFLong outperforms traditional BCF, significantly improving

predictive accuracy and treatment effect estimation in the presence of temporal correlation and individual-level variability, while remaining robust to sparsity in the random effects. We then showcased our model on the NO.MS dataset. Here, BCFLong significantly improved predictive performance and effectively captured clinically meaningful longitudinal patterns in brain volume change, which would have otherwise remained undetected, demonstrating the importance of accounting for within-individual correlations.

These findings demonstrate BCFLong's ability to enhance treatment effect estimation and outcome modelling in longitudinal large-scale studies. By integrating random effects with a sparsity-inducing prior, BCFLong provides a robust and interpretable framework for analysis of longitudinal data, with applications to healthcare and beyond.

[1] Ann-Marie Mallon et al. (2021). Advancing data science in drug development through an innovative computational framework for data sharing and statistical analysis. *BMC Medical Research Methodology* 21, 1–11.

[2] Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1), 266–298.

3: A Bayesian Framework for Measuring the Information Cost of Interim Decisions in Group Sequential Trials

Gianmarco Caruso¹, William F. Rosenberger², Pavel Mozgunov¹, Nancy Flournoy³

¹MRC Biostatistics Unit, University of Cambridge, Cambridge, United Kingdom

²Department of Statistics, George Mason University, Fairfax, VA, USA

³Department of Statistics, University of Missouri, Columbia, MO, USA

Group sequential designs are increasingly adopted in clinical research to enable interim analyses and potential early stopping for efficacy or lack of benefit. While these adaptations improve trial efficiency and ethical considerations, they introduce a selection bias that alters the final inference: by adopting the Bayesian perspective, Flournoy and Tarima (2023) show how failing to account for interim decisions in the model results in biased inference and credible intervals that are overly optimistic in the direction of the decision made. Conversely, properly accounting for interim decisions in the likelihood model reflects the additional uncertainty in the posterior estimates. Drawing on information theory, we adopt a Bayesian entropy-based measure to quantify this cost of adaptation at each interim phase, and show how this can be used for post-hoc evaluation of interim decisions, with a particular focus on the trial's final decision. Similarly to what Tarima and Flournoy (2024) found using

Fisher information, we find that the closer the observed statistic to the decision boundary, the higher this cost of adaptation, reflecting the limited evidence supporting the decision made. We illustrate the use of the proposed measure in a retrospective evaluation of a multi-stage group sequential trial. By comparing alternative decision boundaries and prior specifications, we show how this measure can enhance the understanding of trial results and inform the design of future adaptive studies. Finally, we present an expected version of this metric to guide clinicians in choosing decision boundaries in a pre-experimental phase. This guidance may complement traditional strategies based on type-I error control, such as those provided by Pocock or O'Brien-Fleming boundaries, by offering insights into the potential loss of information on the treatment effect at each interim phase.

Flournoy, N., & Tarima, S. (2023). Posterior alternatives with informative early stopping. *Statistical Papers*, 64(4), 1329-1341.

Tarima, S., & Flournoy, N. (2024). The cost of sequential adaptation and the lower bound for mean squared error. *Statistical Papers*, 65(9), 5529-5553.

4: Bayesian Semiparametric Modelling of Biomarker Variability in Joint Models for Longitudinal and Survival Data

Sida Chen, Marco Palma, Barrett Jessica

University of Cambridge, United Kingdom

Background In clinical and epidemiological studies, there is growing interest in examining the role of within-individual variability (WIV) patterns in longitudinal biomarker data, as emerging evidence suggests that WIV may offer valuable insights and improve predictive power for disease risk and progression. Joint models for longitudinal and time-to-event data (JM) provide a statistically rigorous framework for inferring potential associations between biomarker WIV and clinical outcomes and performing dynamic risk predictions informed by WIV. However, WIV measures themselves can be challenging to estimate reliably, and inferential results can be sensitive to model setting. A motivating example arises when WIV is characterized by biological variability in terms of curvature or wigginess patterns in the underlying biomarker trajectory [1,2]. Existing findings underscore the need for further research.

Methods Motivated by Wang et al. [1,2], we investigated novel modelling strategies for trajectory-based biomarker WIV within the JM context. We propose the use of two state-of-the-art semiparametric approaches to model biomarker WIV: one based on penalized orthogonal P-splines and the other on functional principal component analysis (FPCA). We

compare them with a standard spline-based approach used in Wang et al. [1,2]. For all approaches, we formulated the JM within a Bayesian framework due to its advantages in modelling and computation. Using a Monte Carlo simulation study, we empirically assessed the estimation performance of these approaches under various model settings.

Results In general, the association of trajectory-based WIV with a time-to-event outcome is challenging to estimate robustly under realistic data settings. Among the three approaches, the standard method is the least robust and suffers from convergence issues as models become more complex, while FPCA tends to provide less bias for most parameters across scenarios. When the focus is on prediction, the FPCA-based approach demonstrates the best overall performance, achieving predictive accuracy close to that of the true model across all scenarios. Results from a real-data comparative analysis using data from the Danish cystic fibrosis registry are also anticipated.

Conclusion Estimating the association of trajectory-based WIV with a time-to-event outcome is challenging and requires careful model consideration and interpretation of results. Overall, FPCA appears to be a promising approach for use in JM, particularly when the focus is on prediction.

References [1] Wang et al. *Biostatistics* 2024

[2] Wang et al. *Ann. Appl. Stat.* 2024

5: Bayesian Analysis of the Causal Reference-Based Model for Missing Data in Clinical Trials

Brendah Nansereko¹, Marcel Wolbers², James Carpenter³, Jonathan Bartlett⁴

¹London School of Hygiene and Tropical Medicine, United Kingdom

²Data and Statistical Sciences, Pharma Development, Roche, Basel, Switzerland

³London School of Hygiene and Tropical Medicine, United Kingdom

⁴London School of Hygiene and Tropical Medicine, United Kingdom

Reference based imputation (RBI) methods, proposed by Carpenter et al. (2013), are widely used to handle missing data after the occurrence of intercurrent events (ICEs) in randomized clinical trials. These methods assume no data collection after the ICE. Conventionally, the variance for reference-based estimators was obtained using Rubin's rules but this is biased compared to the repeated sampling variance of the point estimator, due to uncongeniality. Repeated sampling variance estimators were proposed as an alternative to variance estimation

for reference-based estimators. However, these have a property that they decrease as the proportion of ICEs increases. Currently, no frequentist or Bayesian framework method has been developed under which Rubin's variance estimator provides correct inference.

White et al. (2019) introduced a causal model incorporating the concept of a 'maintained treatment effect' post-ICE, showing that reference-based estimators are special cases within this framework. Building on this framework, we propose using a prior distribution for the maintained effect parameter to account for uncertainty about the reference-based assumption using the Bayesian framework. The proposed Bayesian causal model (BCM) provides inference for reference-based estimators that explicitly reflects our uncertainty about how much treatment effects are maintained following the occurrence of ICEs.

A simulation study compared the BCM with existing RBI methods. The analysis used 5000 simulations, incorporating the BCM with fixed and prior distributions on the maintained effect parameter k_0 . Results showed that incorporating a prior distribution on the maintained treatment effect parameter increased posterior variance, particularly in high ICE scenarios, reflecting the impact of the greater uncertainty about the reference-based assumptions. When uncertainty in k_0 was introduced, posterior standard deviation increased with higher ICE rates, aligning with the principle that treatment effect uncertainty should grow as missing data proportions rise.

Application of the method requires pre-specification of the prior distribution for the maintained treatment effect. This mandates careful clinical considerations about the likely impact of ICEs on post-ICE outcomes. The approach can also be used for sensitivity analyses, enabling assessment across varying prior assumptions.

J. R. Carpenter, J. H. Roger, and M. G. Kenward. Analysis of longitudinal trials with protocol deviation: a framework for relevant, accessible assumptions, and inference via multiple imputation. *J Biopharm Stat*, 23(6):1352–71, 2013. doi: 10.1080/10543406.2013.834911.

Ian White, Royes Joseph, and Nicky Best. A causal modelling framework for referencebased imputation and tipping point analysis in clinical trials with quantitative outcome. *Journal of Biopharmaceutical Statistics*, 30(2):334–350, 2020. doi: 10.1080/10543406.2019.1684308.

Multi-Omics Data Integration

Monday, 2025-08-25 16:00 - 17:30, ETH E23

Chair: Charlotte Soneson

1: Unsupervised Factor-Based Methods for Multi-Omics Data Integration

Bernard Isekah Osang'ir^{1,2}, Jürgen Claesen^{2,3}, Ziv Shkedy², Surya Gupta¹

¹Belgian Nuclear Research Centre (SCK • CEN), Mol, Belgium

²I-Biostat, Hasselt University, Diepenbeek, Hasselt, Belgium

³Department of Epidemiology and Data Science, Amsterdam UMC, Amsterdam, The Netherlands

Background Multi-omics data integration is essential for advancing precision medicine and systems biology, allowing for a holistic understanding of complex biological processes. Unsupervised factor-based methods offer a powerful approach for identifying latent patterns within high-dimensional and heterogeneous biomolecular datasets. However, the performance of these methods in integrating diverse omics modalities remains underexplored. This study systematically benchmarks three widely used factor-based methods—Multi-Omics Factor Analysis (MOFA), Multiple Factor Analysis (MFA), Group Factor Analysis (GFA), and proposed a new method in the context of multi-omics data integration, the Factor Analysis for Bicluster Acquisition (FABIA).

Methods We evaluated these methods using real-world and simulated multi-omics datasets. The real datasets included (1) a Chronic Lymphocytic Leukemia (CLL) study integrating DNA methylation and drug response profiles and (2) an experimental radiation dataset capturing transcriptomic and proteomic data from irradiated mouse brain tissues. Using the R package SUMO, we simulated datasets with predefined latent factors, varying noise levels, and different data distributions. Each method was assessed for its ability to (i) recover predefined latent factors, (ii) capture shared and unique variance components, and (iii) maintain robustness across noise conditions. Performance was evaluated using Jaccard index, Pearson correlation of factor scores, feature weights, and metrics such as sensitivity, specificity, and accuracy. Additionally, we explored semi-supervised integration, where the FABIA method is used for guided feature selection enhanced signal discovery.

Results All four methods successfully identified latent biological patterns, however, the MOFA and the FABIA methods showed the strongest performance, capturing similar biological signals with high agreement. FABIA excelled at low-to-moderate noise but declined

under high noise, while MOFA remained stable. The MFA method seems to be robust but struggled with extreme noise, and the GFA method was the most affected. Semi-supervised learning enhanced variance explanation and feature selection, refining latent factor detection and improving integrative multi-omics analysis, particularly in complex and noisy datasets.

Conclusions Factor-based methods are crucial for multi-omics integration, providing scalable and interpretable solutions for extracting biological insights. The benchmark analysis of these methods, highlights their strengths and limitations in handling noise and dataset complexity. Multi-omics plays a key role in understanding disease mechanisms, identifying biomarkers, and advancing precision medicine. As the availability of multi-omics data grow, robust statistical frameworks for multi-omics data integration is essential for uncovering complex biological relationships and biomedical applications.

2: Multi-Modal Integration Reveals the Joint Role of Electrocardiogram, Imaging and Genetics in Cardiovascular Risk

Andrea Mario Vergani^{1,2}, Francesca Ieva^{1,2}, Marco Masseroli¹, Emanuele Di Angelantonio^{2,3}

¹Politecnico di Milano, Milano, Italy

²Human Technopole, Milano, Italy

³University of Cambridge, Cambridge, United Kingdom

Background / Introduction The recent availability of biobank-scale multi-modal health-care data offers invaluable opportunities for personalised risk profiling and studying omics' biological impacts on disease. In the cardiovascular field, multi-modal data is largely collected, including genetics, electrocardiogram (ECG) and imaging; however, the biological interactions of these complex modalities and their impacts across disease subtypes are still unclear. For this reason, this study leverages omics fusion and survival analysis to explore the value that imaging, ECG and genetics can provide to cardiovascular risk prediction, as well as their interplay when integrated together.

Methods Analysing data from the UK Biobank, including a panel of Polygenic Risk Scores (PRSs) available for 485,000 participants, ECG-derived measures (e.g., QRS duration) from 40,000 individuals, and Cardiac Magnetic Resonance (CMR) measures (e.g., left ventricular ejection fraction) from 30,000 subjects, we leveraged a state-of-the-art omics fusion method - Multi-Omics Factor Analysis - to integrate such modalities into a joint low-dimensional feature space. The embeddings were then fitted in survival analysis studies on about 20,000 subjects healthy at baseline with PRS, ECG and CMR data available, targeting various cardiovascular disease subtypes (e.g., coronary artery disease). Moreover, we analysed the

integrated feature space to evaluate the different impacts of PRSs, ECG and CMR on our embeddings and, in turn, on disease risk.

Results When integrating the modalities with Multi-Omics Factor Analysis, we observed a clear interplay between genetic, ECG and CMR information, especially in the first principal embeddings: as an example, the first integrated factor was mostly ECG-driven, but also explained part of the variance of PRS and CMR datasets; the second one, instead, mostly captured CMR measures, but also retained cardiovascular-related genetic information. Overall, our latent representation integrated omics interactions in a 10-dimensional space, explaining nearly 50% of the variance of the CMR dataset and about 30% of the PRS and the ECG ones each. The three modalities together proved to predict time-to-cardiovascular event, achieving a mean cross-validation concordance index of 0.71; specifically, the first and second joint factors had statistically significant effects on cardiovascular risk prediction, together with the fourth one, which captured exclusively genetic variability.

Conclusion We propose a novel approach employing population-level multi-omics fusion to integrate ECG, CMR and genetic measures for better-informed cardiovascular risk prediction. The integrated complex modalities improved prognostic disease prediction and demonstrated statistical significance and value beyond traditional clinical covariates, thus contributing to enhanced personalisation for cardiovascular risk stratification.

3: Analyzing Protein Folding Dynamics Using Multi-Dimensional Varying Coefficient Models

Jürgen Claesen

Amsterdam UMC, Netherlands, The

Proteins are initially synthesized as unstructured polymers on ribosomes and fold into functional forms. The folding process is influenced by both intrinsic protein properties and interactions with external factors, leading to a range of behaviors from rapid folding to stable unfolding. Protein folding mechanisms can be investigated using pulsed hydrogen-deuterium exchange mass spectrometry (HDX-MS), which tracks both local and global exchanges. In a pulsed HDX experiment, proteins are exposed to deuterium for a set period, causing hydrogen atoms to exchange with deuterium and resulting in a measurable mass increase. This change in mass is captured by a mass spectrometer coupled with liquid chromatography (LC). By monitoring shifts in retention time and mass at various stages of the folding process, the mechanism and rate of folding can be determined.

To examine folding differences across multiple proteins, we developed a multi-dimensional varying coefficient model. In this model, protein mass and retention time are treated as covariates within one-dimensional smooth functions. The product of these one-dimensional smooths creates a smoothed surface. We also incorporated interactions between categorical variables (protein identity and folding time) and the two smooths (mass and retention time), yielding multiple smoothed surfaces. The model, structured with main effects and interactions, allows for the estimation and testing of smoothed differences between the reference surface and other surfaces, with results evaluated using simultaneous confidence bands.

4: X-Med: Cross-Modal Integration for Sequential Intelligence in Aftercare of Kidney Transplant Recipients

Aditya Kumar¹, Simon Rauch², Mario Cypko¹, Oliver Amft^{1,2}

¹Hahn-Schickard, Freiburg, Germany

²Intelligent Embedded Systems Lab, University of Freiburg, Germany

Introduction Early prediction of outcomes for kidney transplant recipients (KTRs), including graft loss and rejection, is crucial for improving post-transplant care. Leveraging multimodal data offers complementary insights, including structured (e.g., demographics, laboratory results, vitals) and unstructured (e.g., clinical notes) information from Electronic Health Records (EHRs). Structured longitudinal data are often affected by missing values, irregular sampling and asynchronous measurements. Unstructured text demands context-aware processing to extract relevant clinical information. To address the aforementioned data challenges, we propose a unified patient representation learning approach that models each data modality individually and integrates them into a shared embedding space. The learned patient-level embeddings are evaluated on downstream prediction tasks and provide interpretable, disentangled features representation.

Methods Structured data, including both time-varying and static features, are modelled using a Time-Aware LSTM with self-attention. Static features are reintroduced at each time step to account for their influence overtime. The Time-Aware LSTM architecture addresses temporal dependencies, irregular sampling, asynchronous features, and missing data. Unstructured clinical notes are embedded using a pretrained sentence transformer (gte-large), finetuned on clinical texts. Representations from both modalities are fused via cross-attention into a shared patient embedding space. The training process enforces a disentangled embedding space. The embeddings are evaluated on outcome prediction tasks (graft loss, rejection, and mortality) using the NephroCAGE dataset [1]. Interpretability is assessed with SHAP values, assessing whether the influential features align with medical

domain knowledge. Additionally, we use the mutual information gap (MIG) and separate attribute predictability (SAP) to quantify disentanglement.

Results Preliminary results demonstrate that our model achieves state-of-the-art performance in predicting graft loss and rejection (ROC-AUC = 0.95 and 0.81). SHAP analyses indicate that the model captures established risk factors for kidney patients. Ablation studies confirm that complementary information from different modalities is effectively integrated. Ongoing work focuses on evaluating disentanglement using MIG and SAP to optimise training settings without compromising predictive performance.

Conclusion Our results highlight the potential of multimodal patient representations for outcome prediction in kidney transplantation. Integrating structured and unstructured data improves performance, yet balancing predictive power and disentanglement remains challenging. Future work will explore training strategies to enhance latent space interpretability while preserving clinical relevance and accuracy.

Reference [1]. Schapranow, Matthieu-P., et al. "NephroCAGE—German-Canadian Consortium on AI for Improved Kidney Transplantation Outcome: Protocol for an Algorithm Development and Validation Study." *JMIR Research Protocols* 12.1 (2023): e48892.

5: Modeling Interdependencies in Multiomic Spatial Analysis

Giulia Capitoli¹, Vanna Denti¹, Veronica Vinciotti², Ernst Wit³

¹University of Milano-Bicocca, Italy

²University of Trento, Trento

³University of Svizzera Italiana, Switzerland

Introduction Understanding the dependency structure among a large number of molecules is a central goal in biology, particularly in the context of disease research and biomarker discovery. However, real-world data often present significant challenges due to their heterogeneous nature. Samples are frequently collected under varying spatial and temporal conditions, leading to differences in network structures across groups. In such cases, the assumption of independent and identically distributed (i.i.d.) data becomes unrealistic. Applying a single graphical model to the entire dataset risks overlooking meaningful group-specific variations, while fitting separate models for each group fails to leverage shared patterns between groups and often requires pre-labeled group information.

Methods To address these challenges, Gaussian Graphical Mixture Models (GGMMs) have

emerged as a promising solution. GGMMs assume that data arise from a mixture of Gaussian distributions, where each component represents a subgroup with its unique network structure. This framework enables the simultaneous identification of cluster memberships and the modeling of intra-cluster dependencies, offering a principled approach for analyzing heterogeneous and high-dimensional data.

Central to the proposal is the extension of Gaussian Graphical Mixture Models (GGMMs) to incorporate spatial dependencies and multimodal data, resulting in a Spatial Gaussian Copula Graphical Mixture Model (SGCGMM).

Results By leveraging techniques such as markov random fields, sparse precision matrices, and copula models, the model achieves interpretable and scalable results, accounting for spatial correlations, integrate diverse molecular profiles, and handle mixed data types and overcoming current limitations in handling high-dimensional, nested, and noisy data.

The methodology will be applied to mass spectrometry imaging (MALDI-MSI) datasets, which provide spatially resolved molecular profiles on the same biopsy tissue section. These efforts will characterize interdependencies between molecular families, identify spatial biomarkers, and provide insights into tumor microenvironments and disease mechanisms.

Conclusion Beyond the motivating application, the tools and theoretical advancements developed in this work will have broader applicability, driving progress in multiomic data integration and precision medicine. A computational pipeline and user-friendly tools are under development to ensure the accessibility and scalability of the proposed approaches.

Biomarker Studies & Diagnostic Tests

Monday, 2025-08-25 16:00 - 17:30, ETH E21

Chair: Annette Kopp-Schneider

1: Assessing Diagnostic Accuracy for Three-Class Classification Problems

Maria C. Pardo, Alba M. Franco-Pereira, Victor M. Sierra

Complutense University of Madrid, Spain

Keywords: diagnostic accuracy, power, volume under the ROC surface (VUS), overlap measure (OVL)

Background / Introduction

Diagnostic testing is an extremely important aspect of medical care. In many situations the diagnostic decisions are not always binary. An early or intermediate disease stage usually occurs as individuals transition from the healthy stage to the fully diseased stage. To summarize a diagnostic test's overall ability to simultaneously discriminate three diagnostic groups, the volume under the curve (VUS) is one of the most well-known measures which generalizes the notion of the area under the curve (AUC) in the two-class problem.

Methods

However, VUS is limited in their ability to fully capture the complexities of some scenarios in the three -class problem as well as AUC in the two-class problem. Pardo and Franco (2025) explored the advantages of the Overlap measures (OVL) over the AUC to assess the accuracy of a medical diagnostic test in the binary case. In this work, extension of this measure has been studied for three-class classification problems. We study methods for estimating OVL for three groups under both parametric and non-parametric frameworks. Furthermore, we propose a testing process for its statistical significance.

Results

The size and power of the proposed methods for testing the utility of biomarkers drawn from normal, lognormal gamma distributions and mixture of them at different sample sizes are evaluated. Most of cases, our proposal is preferred to VUS.

Conclusion

In some situations, VUS tends to perform very poorly, with power values approaching to 0.05. Therefore, VUS would naively lead to rejection of informative biomarkers. However, OVL outperforms VUS in these situations, making it a valuable tool in the biomarker field.

References Pardo, M.C. and Franco-Pereira, A.M. (2025). Overlap measures against ROC summary indices. *Statistical Science*, in press

2: Biological Age Estimation in the Estonian Biobank Based on NMR Metabolomics Data and Phenotype

Mara Delesa-Velina¹, Krista Fischer^{1,2}, Estonian Biobank Research Team²

¹Institute of Mathematics and Statistics, University of Tartu, Estonia

²Institute of Genomics, University of Tartu, Estonia

Background With aging populations on the rise, there is increasing interest in studying biological age measures. NMR metabolites, small molecules involved in metabolic pathways detected using NMR spectroscopy, have shown promise in estimating biological age. With now more than 200,000 participants in the Estonian Biobank having NMR blood metabolite data available, we aim to develop a model for predicting all-cause mortality and estimating biological age.

A common approach for biological age estimation is regression modelling, where age is used as the dependent variable. This approach produces biological age estimators (aging clocks) that predict an individual's age as precisely as possible. However, this does not imply that individuals with biological age estimate exceeding their chronological age have a higher risk of disease or a shorter lifespan. An alternative approach is to define biological age so that it is directly related to the underlying risk level. We propose such an approach based on a parametric survival model.

Methods We develop the model using the first cohort of biobank participants ($n=31,359$, recruited between 2002 and 2010, mean follow-up 13.3 years, SD 4.4 years). We validate the model using the second cohort of the biobank ($n=118,664$, recruited from 2018 onwards, mean follow-up 5.2 years, SD 0.7 years).

We employ a Cox proportional hazards model with age as a timescale and stepwise selection to identify NMR metabolite biomarkers independently associated with 10-year mortality. We model an individual's survival probability using a parametric Gompertz distribution with NMR score, prevalent disease, and phenotype as covariates. Finally, we define survival-based biological age as the age where the individual's current survival probability, given their covariate profile, equals the survival probability of an average individual in the cohort. We estimate biological age acceleration (BAA) as the difference between biological and chronological age.

Results The NMR score comprises 17 metabolic biomarkers and is highly associated with mortality in both the development and validation cohorts, with HR (per SD of NMR score) of 1.78 (95% CI 1.73–1.83) and 1.79 (95% CI 1.74–1.84), respectively. The survival-based biological age estimate is symmetrically distributed around the chronological age. BAA estimate is a powerful predictor of 5-year survival (C-index 0.762, Cox model with age timescale) in the validation cohort and remains informative for the age group over 70 (C-index 0.673).

Conclusion Survival-based biological age estimate based on NMR metabolite score is a more powerful predictor of mortality than the chronological age adjusted by common phenotypic predictors.

3: Bézier Curve Parametric Method for Approximating ROC Curves in the Context of Multiple Clinical Decision Thresholds

Denys Prociuk¹, Brendan Delaney¹, Francesca Fiorentino^{1,2}

¹Imperial College London, United Kingdom

²University of Leeds, United Kingdom

Background / Introduction Receiver Operating Characteristic (ROC) curves are fundamental in clinical decision-making and are widely used to assess diagnostic test performance. However, selecting cut-off points to guide a clinical decision can be challenging. Traditional approaches—such as Youden's J statistic or clinician judgment—often have limitations, especially when multiple thresholds would have better clinical utility. We propose using the Bézier curve parametric method to fit a curve to diagnostic test data and to determine cut-off

points by leveraging on the fitted curve's shape and its rate of change.

Methods We use the RECAP-V1[1] study data to demonstrate the application of the Bézier curve method. RECAP-V1 produced a ROC curve to determine which patients were at high risk of hospitalisation for COVID-19.

Using the non-linear least squares methods, we identified "control" points for optimising the fitting of both cubic and quadratic Bézier curves. These control points were then used to identify candidate cut-off points, which were compared to thresholds derived from expert clinician judgment in RECAP-V1 (Green/Amber-Amber/Red thresholds for risk). Additionally, we examined the Bézier curve's curvature as an alternative strategy for identifying a single optimal cut-off to compare the use of Bézier to Youden's method. Sensitivity, specificity, and interval likelihood ratios (ILR) were used as performance metrics.

Results The quadratic Bézier approach yielded a Green/Amber threshold with 91% sensitivity (ILR of 0.27), and an Amber/Red threshold with 97% specificity (ILR of 6.66). The cubic method produced similar outcomes, demonstrating the robustness of the approach. When comparing to expert clinical judgement, the Green/Amber threshold showed a similar sensitivity (91% vs. 90%, ILR 0.27 vs 0.16), while the Amber/Red threshold demonstrated higher specificity (97% vs. 90%, ILR 6.66 vs. 6.00). Hence, Bézier-derived thresholds were in close agreement with those selected by clinicians. Curvature-based analysis provided an alternative single cut-off point that closely matched Youden's J statistic.

Conclusion Bézier curve fitting offers a robust method for selecting ROC curve cut-off points, aligning closely with expert clinical judgment. It can aid non-experts in the identification of multiple thresholds for clinical decision. For RECAP-V1 it improved sensitivity and specificity hence could have improved clinical decision-making. Future research should explore its applicability across a range of diagnostic models.

[1] Espinosa-Gonzalez et al. Remote COVID-19 Assessment in Primary Care (RECAP) risk prediction tool: derivation and real-world validation studies. Lancet Digital Health. 2022;4(9):e646–e656. <https://doi.org/10.1016/j.diatho.2022.100123>

4: Deriving Cost-Effective Neyman-Pearson Classifier with Multiple-Modality Detection Tools

jiaming Qiu, Yingqi Zhao, Yingye Zheng

Fred Hutchinson Cancer Center, United States of America

Background

In binary medical decision-making, such as early disease detection, the goal is to identify patients at risk for malignant outcomes while avoiding unnecessary invasive diagnostic procedures. Neyman-Pearson (NP) classifiers are commonly used to control the false positive rate (FPR) within an acceptable threshold while maximizing the true positive rate (TPR). However, when multiple testing modalities are available, there is a challenge in balancing the benefits of comprehensive disease detection with the costs and complications of extensive testing. In prostate cancer diagnosis, for example, a biopsy is an invasive procedure that may not be necessary for many low-risk patients. Current clinical guidelines suggest the use of various biomarker tests and multiparametric magnetic resonance imaging (mpMRI) for risk stratification, yet clinicians face uncertainty in selecting and sequencing these tests.

Methods:

We propose a sequential decision-making framework within the NP classifier paradigm to address these challenges in prostate cancer diagnosis. Specifically, we develop a 2-step diagnostic protocol that utilizes biomarker tests and MRI results. In the first step, patients are categorized based on biomarker test values: those with values below a low threshold are sent home without further testing, those exceeding a high threshold are referred for a biopsy, and those with intermediate values undergo MRI for additional information. In the second step, the biomarker and MRI results are combined to decide whether a biopsy is necessary for patients who underwent MRI. The objective is to minimize unnecessary biopsies, maintain an acceptable false negative rate for aggressive cancers, and reduce procedural costs by limiting the number of patients who undergo further testing. **Results:**

The proposed sequential rule effectively minimizes false positives (unnecessary biopsies) while keeping the false negative rate for aggressive cancers within clinically acceptable limits. By selectively using tests, it reduces the number of patients who undergo subsequent procedures, leading to a reduction in overall procedural costs without compromising diagnostic accuracy. The trade-offs between limiting initial testing and controlling the false positive rate are quantified, optimizing the balance between diagnostic performance and cost efficiency. **Conclusion:**

The sequential decision-making protocol presented in this study offers a more cost-effective and personalized approach to prostate cancer diagnosis. By optimizing the use of biomarker tests and MRI, the method minimizes unnecessary procedures and maximizes diagnostic accuracy, providing a framework that can be adapted for other medical decision-making contexts involving multiple tests.

5: Improving Biomarker Diagnostic Accuracy with the Likelihood Ratio Transformation

Ainesh Sewak¹, Vanda Inacio²

¹University of Bern, Switzerland

²University of Edinburgh, Scotland

Introduction Accurate biomarker-based diagnostic and screening tests rely on the receiver operating characteristic (ROC) curve to assess classification performance. However, in some cases, the empirical ROC curve of a biomarker is improper or 'hooked', meaning it crosses below the diagonal and fails to provide a reliable decision rule. It has been established since the invention of ROC curves that mapping biomarkers to the likelihood ratio scale yields a mathematically optimal decision rule and ensures a proper ROC curve. However, despite its theoretical appeal, there is surprisingly little literature on this approach, with only a few parametric developments addressing it.

Methods We present three models for transforming biomarkers to the likelihood ratio scale, each leading to an optimal decision rule. The parametric binormal model provides a closed-form transformation under Gaussian assumptions. This serves as a foundation for more flexible approaches. Next, the semiparametric approach leverages flexible distributional regression models, allowing for marginal density estimation without strict parametric constraints. Finally, we demonstrate how additive logistic regression can achieve the same transformation using standard binary regression techniques. For each method, we establish theoretical properties that ensure proper ROC curves and optimal classification performance.

Results Through simulations and analysis of three biomarker datasets, we demonstrate that transforming improper biomarkers to the likelihood ratio scale consistently improves diagnostic accuracy. The improvement is most pronounced when the original ROC curve is highly improper, while for already proper ROC curves, the transformation has minimal effect.

Conclusion The likelihood ratio transformation offers a simple and powerful solution for correcting improper ROC curves and improving biomarker diagnostic accuracy. Our results indicate that transforming to the likelihood ratio scale should be the default, especially when biomarkers exhibit impropriety. This work has broad practical implications for clinical biomarker evaluation and its implementation is readily accessible using standard statistical software. *Keywords:* ROC curve, likelihood ratio, biomarkers, regression, generalized additive models.

Clinical Trials with Longitudinal and Clustered Data

Tuesday, 2025-08-26 09:15 - 10:45, Biozentrum U1.131

Chair: Garth Tarr

1: Location-Scale Latent Process Model for Repeated Ordinal Patient Reported Outcomes

Agnieszka Król¹, Robert Palmér², Jacob Leander², Cécile Proust-Lima³, Alexandra Jauhiainen⁴

¹R&I Biometrics and Statistical Innovation, Late R&I, BioPharmaceuticals R&D, AstraZeneca, Warsaw, Poland

²Clinical Pharmacology and Quantitative Pharmacology, Clinical Pharmacology & Safety Sciences, R&D, AstraZeneca, Gothenburg, Sweden

³University of Bordeaux, Inserm, Bordeaux Population Health Research Center, UMR 1219, Bordeaux, France

⁴R&I Biometrics and Statistical Innovation, Late R&I, BioPharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden

Patient reported outcomes (PROs) are collected on a daily basis in clinical trials to measure patients' quality of life, e.g. symptoms. Often these data are reported in a small-range ordinal scale and analyzed without considering their longitudinal aspect. The emergence of electronic data collection methods for home-based measurements has enabled routine, daily capture of various symptom scores, highlighting the need to develop statistical methods for the analysis of these intensive ordinal longitudinal data. Of interest are both their mean structure over time and variability which is known to be linked to disease progression and to be affected by treatments. To model the dynamics of ordinal PROs, we propose a location-scale latent process model which includes two types of variability across patients: individual underlying level flexibly modelled over time (e.g., with splines) using random effects and covariates, and individual short-term variability with a variance of the error which is expressed as a linear structure of covariates such as treatment arm and a patient-specific random intercept. The model is estimated in a Maximum Likelihood framework with an interface in R. The high-dimensional intractable integrals in the optimization are approximated using the Quasi-Monte Carlo method. The estimation procedure is validated by a simulation study, and we apply the methodology to data from two clinical trials, one in asthma and one in COPD, to evaluate the effect of treatment on dynamics of the respiratory symptoms and their variability.

2: Causal Approaches for the Design and Long-Term Treatment Effect Estimations of Hybrid Randomized Clinical Trials with Longitudinal Outcomes

Xiner Zhou¹, Herbert Pang², Christiana Drake¹, Hans Ulrich Burger³, Jiawen Zhu²

¹University of California, Davis

²Genentech

³Roche

Background / Introduction Incorporating external data, such as external controls, holds the promise of improving the efficiency of traditional randomized controlled trials especially when treating rare diseases or diseases with unmet needs.

Methods In the first part of the talk, we describe novel weighting estimators grounded in causal inference. From a trial design perspective, operating characteristics including Type I error and power are particularly important and results will be presented. In the latter part, we describe proper estimation and inference of long-term treatment effect during the open-label extension phase in the absence of placebo-controlled patients. Within the framework of causal inference, we propose several difference-in-differences type methods and a synthetic control method for the combination of randomized controlled trials and external controls.

Results For the first part, our proposed weighting estimators achieve significant power gain, while maintaining Type I error close to the nominal value of 0.05 under the simulation settings investigated. For the latter part, our realistic simulation studies demonstrate the desirable performance of the proposed estimators in a variety of practical scenarios. In particular, difference-in-differences type methods outperform synthetic control method and are the recommended methods of choice in scenarios similar to the ones we have investigated.

Conclusion In both studies, we assessed in our realistic simulation studies representing a variety of practical scenarios and provided an application through a phase III clinical trial in rare disease.

References Zhou, X., Pang, H., Drake, C., & Zhu, J. (2024). Causal estimators for incorporating external controls in randomized trials with longitudinal outcomes. *Journal of the Royal Statistical Society Series A*, qnae075.

Zhou, X., Pang, H., Drake, C., Burger, H. U., & Zhu, J. (2024). Estimating treatment effect in randomized trial after control to treatment crossover using external controls. *Journal of biopharmaceutical statistics*, 34(6), 893–921.

3: Prediction Powered Inference for Trials with Survival Outcomes

Maylis Tran¹, Pierre-Emmanuel Poulet¹, Bruno Jedynak², Sophie Tezenas du Montcel¹

¹Sorbonne University, Paris Brain Institute, INSERM, CNRS, INRIA, Assistance Publique-Hôpitaux de Paris (APHP), University Hospital Pitié-Salpêtrière, Paris, France.

²Department of Mathematics and Statistics, Portland State University, 1855 SW Broadway, Portland, 97201, Oregon, USA

Background Recruiting patients for clinical trials in rare neurodegenerative diseases is challenging, because of ethical concerns surrounding placebo enrollment and difficulties in determining optimal sample size for statistical significance. Innovative statistical techniques such as prediction powered inference for clinical trials (PPCT) (Poulet et al., 2025) can improve trial statistical power and reduce sample size requirements. Amyotrophic Lateral Sclerosis (ALS) clinical trials are driven by two key outcomes: survival as the primary measure and the ALSFRS-R score as the secondary measure. This study aims to apply PPCT to estimate the average treatment effect, first by using the classical PPCT estimator for continuous outcomes and then by adapting the method for survival outcomes.

Methods Natural history data from various observational datasets were used to train a progression model capturing the natural disease development in untreated patients (digital placebo twins). A joint model was applied to account for ALS disease progression features, which include both longitudinal and event-based data (Disease Course Mapping). Then, we applied this model to predict the disease progression of clinical trial patients. The PPCT method incorporated these predictions into trial analyses. PPCT first compares treated patients with their digital placebo twins, to estimate the average treatment effect (continuous outcome) or the hazard rate estimator (survival outcomes). To ensure robustness, it accounts for prediction errors and placebo effects by comparing the predicted placebo progression to the observed one, thereby debiasing the final estimations. We adapted PPCT to survival outcomes, by applying the general PPCT estimator formula to the hazard rate estimator.

Results By applying PPCT to the Kaplan-Meier estimator, we enhanced the statistical power of the log-rank test. We further validated the methodology by using PPCT on the treatment effect estimator, which compares the mean ALSFRS-R score progression between the placebo and treatment groups. We observed a significant reduction in its variance with narrower confidence intervals. To assess the accuracy of our predictions, we analyzed the correlation between the predicted and observed progression (R^2). Higher R^2 values were associated with narrower confidence intervals.

Discussion Our study underscores the importance of leveraging high-quality observational data to accurately train the joint spatiotemporal model and reach high statistical power

for both average treatment effect estimation and log-rank test. The application of PPCT methodology and its adaptation to survival outcomes is a relevant method either to reduce the sample size requirement (pre-trial) or to improve the statistical analysis of the clinical trial results (post-trial).

4: Repeated Measurements Modelling of Titration Effects in Multi-Arm Clinical Trials

Emma Ove Dahl^{1,2}, Philip Hougaard¹

¹Lundbeck, Denmark

²Danish Cancer Institute, Denmark

Background Traditionally, a clinical trial with multiple treatment arms is analysed as if the arms are completely unrelated to each other. However, the trial design may lead to some arms having shared features. A common example is the study of several doses of the same drug, where the trial has a titration phase, so that the doses are the same at the first few visits. The idea of the proposal is to account for this feature in the analysis, following the basic principle of analysing a trial according to its design.

Methods Technically, it is very easy to account for this feature in the analysis and this makes the results more coherent. The approach is illustrated with results from a depression trial of adjunctive brexpiprazole in doses of 1 and 3 mg/day compared to placebo for patients with inadequate response to antidepressants. The benefits are also documented with simulations, covering both analysis of a full trial and performance in case the trial design calls for an interim analysis.

Results While the precision improvement in placebo comparisons at the final visit is only small, the comparison of the active arms becomes more precise, and it also improves precision of treatment effects at the earlier visits. A further advantage is that interim analyses become more precise because the method better utilizes the measurements at early time points, which are present for patients ongoing at the time of doing the interim analysis. As this is suggested as a drug development approach, we also show how it fits into the estimand framework.

Conclusion Considering the simplicity of the approach, it is beneficial for the primary analysis, even though the precision improvement is only small. The benefits for secondary analyses and interim analyses are more important, substantiating the relevance of the suggestion.

Bayesian Methods 2

Tuesday, 2025-08-26 09:15 - 10:45, Biozentrum U1.141

Chair: Lukas Andreas Widmer

1: Using a Foundation Model for Detecting and Reducing Site-Specific Differences in Federated Meta-Analysis of Regression Models

Patric Tippmann^{1,2}, Max Behrens^{1,2}, Harald Binder^{1,2}

¹Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center – University of Freiburg

²Freiburg Center for Data Analysis, Modeling and AI, University of Freiburg

Multi-centre meta-analysis of regression models can be impacted by discrepancies in relations between covariates across centres. The reasons for such discrepancies are particularly difficult to detect when data cannot be pooled or directly transferred between centres. We propose a two-step, federated approach to (1) detect cross-centre differences and (2) harmonise them under data sharing restrictions. Both steps require a model that captures relations between variables to be trained at a reference centre. The trained model is then transferred to the other centres with restricted data access to diagnose and potentially remove differences. While standard regression models could be used for this task, these might not be flexible enough for reflecting complex patterns. As an alternative, we consider using a foundation model, specifically TabPFN, which encodes a prior distribution on plausible patterns and can be used to obtain a posterior based on the reference centre. A further advantage of this specific approach is robustness to non-linear transformations, outliers and missing data. Furthermore, this flexibility is accompanied by high computational efficiency—requiring only a single forward pass—and suitability for small datasets. Specifically, the trained model is applied locally at each restricted centre to generate predictions for the inaccessible variables. Aggregating these predictions over multiple iterations yields a harmonised estimate of the local correlation structure.

We illustrate these benefits with clinical datasets, with a focus on stroke, demonstrating that a conspicuously different regression coefficient in a restricted collaborating centre shifts substantially toward the reference centre's value after applying our harmonisation approach to the associated correlation structures. Because a collaborating centre retains its raw data locally—receiving only the trained model from the reference centre—the method takes into account regulatory and ethical constraints on data sharing. The resulting harmonised correlations enable more reliable multi-centre analyses while preserving individual-level privacy.

2: Sample Size Calculations for Prediction Model Development: A General Bayesian-Framework using Posterior Distributions to Examine Expected Performance, Degradation and Stability

Richard D Riley¹, Rebecca Whittle¹, Mohsen Sadatsafavi², Glen Martin³, Alexander Pate³, Gary Collins⁴, Joie Ensor¹

¹University of Birmingham, United Kingdom

²The University of British Columbia, Canada

³University of Manchester, United Kingdom

⁴University of Oxford, United Kingdom

Background

For studies developing a clinical prediction model, various sample size calculations exist. However, their underlying theory is often based on standard (unpenalised) regression and extensions to other machine learning approaches are needed.

Objectives

To propose a general Bayesian approach to sample size calculations for model development or updating, based on drawing samples from anticipated posterior distributions, and targeting small reduction in predictive performance ('model degradation') compared to an assumed true model.

Methods

Researchers must provide their candidate predictors, a 'true model' (e.g., regression equation with intercept and predictor effects that match outcome incidence and c-statistic of previous models), and a (synthetic) dataset reflecting the joint distribution of candidate predictors. Then, for a chosen sample size and development strategy, our general approach is fully simulation-based: generate thousands of models and apply each to a large evaluation dataset to produce posterior distributions of individual predictions, model performance and model degradation. However, to substantially improve computational speed for penalised regression (e.g., lasso, ridge), we propose approximating posterior distributions using a one-sample Bayesian analysis that incorporates shrinkage priors alongside the likelihood decomposed into sample size and Fisher's unit information. The derived posterior distributions enable *any* criteria to be examined (e.g., mean and variance of calibration slope; expected degradation in c-statistic; mean width of 95% intervals for individual risk; expected value of sample information for decision-making) to inform the (minimum) sample size required.

Results

We illustrate the approach when developing models in pre-eclampsia and show how they encompass any criteria of existing sample size calculations, whilst additionally allowing researchers to examine variability (instability) of model predictions and degradation in model performance. Focusing on ridge and lasso logistic regression, we demonstrate the Bayesian one-sample analysis via our module *pmssbayes* and compare results/speed with the fully simulation-based approach. We show how the approach informs fairness of models and outline practical options for specifying the 'true model' and case-mix distribution.

Conclusions

Our Bayesian approaches generalise existing sample size proposals for model development, by utilising anticipated posterior distributions conditional on a chosen sample size and development strategy, to inform the sample size required to target appropriate model performance, stability and clinical utility.

3: Efficient Utilization of Dose-Schedule Grids for Optimal Therapeutic Outcomes in Non-Oncology Settings

Lars Andersen, Mitchell Thomann, Thomas Jaki

BOEHRINGER-INGELHEIM PHARMA GMBH & Co KG, Germany

Efficient Utilization of Dose-Schedule Grids for Optimal Therapeutic Outcomes in Non-Oncology Settings

Authors:

Lars Andersen - Boehringer Ingelheim Pharma GmbH & Co. KG,

Mitchell Thomann - Boehringer Ingelheim Pharma GmbH & Co. KG,

Thomas Jaki - University Regensburg

Topics: Bayesian Methods, Simulation Studies, clinical trials - designs

Background Existing dose-schedule finding methods in phase I oncology studies are constrained by toxicity and small sample sizes, resulting in smaller grids and narrow design

spaces. These methods assume predefined dose and schedule ordering and require starting at low doses with structured escalation. In phase II, these assumptions may not hold for efficacy. This simulation study evaluates a more robust modeling framework with separate models for dose and schedule, considering various study designs that allocate across the factorial space with and without optimization criteria as well as with and without an interim analysis.

Methods Across various designs and modelling frameworks, the study measured performance in terms of go/no-go decision, correct minimum effective dose (MED) estimation, mean squared error (MSE), root mean squared error (RMSE), Akaike information criterion (AIC), Bayesian information criterion (BIC) and the number of patients allocated at the true MED.

Results The correct MED estimation is not as robust if there is a drop in sample size and therefore the study being potentially underpowered, nor if there are big variance changes on patient level. Model misspecification worsens performance, decision-making, and estimation. Go/No-Go measurements perform well across designs and scenarios if the true model is selected. MSE and RMSE are robust across scenarios, and patient allocation at the true MED mirrors correct MED estimation. The full factorial design performs well and is robust due to its allocation across the sample space. Adding interim analysis slightly improves performance, especially in patient allocation at the true MED. Optimality designs show minimal differences across scenarios.

Conclusion To conclude, one can say that this simulation study showed a framework for design and modelling in this setting which was developed and tested across a variety of scenarios. In general, this framework was able to show robustness of model selection criteria such as AIC and BIC. Additionally, it also displayed the benefit of an interim analysis to reallocate the patients based on criteria such as the MED.

4: On the Interplay Between Prior Weight and Vague Variance in Robust Mixture Priors

Marco Ratta¹, Gaëlle Saint-Hilary², Pavel Mozgunov³

¹Politecnico di Torino, Italy

²Saryga, France

³Cambridge University, United Kingdom

The use of historical data in complementing current control arm in the context of randomized controlled trials (RCTs) is increasingly attractive, particularly when patient recruitment presents a significant hurdle. This necessitates addressing potential conflicts between histori-

cal and current trial data. Robust Mixture Priors (RMPs) are a prominent dynamic borrowing approach to mitigate this, consisting in combining an historical informative component and a weakly informative robust component via mixture distribution. Once observed the data, the RMP is updated in a posterior distribution that is again a mixture of the individual posterior distributions, with updated posterior weights. The RMP's key feature is its borrowing mechanism, directly proportional to the agreement between historical and current data. High agreement maximizes borrowing; inconsistencies progressively reduce it.

Specifying parameters for normal RMP components, particularly variance of the robust component and mixture weights, presents a challenge, as these parameters significantly influence posterior inferences. Improper normal distributions seem intuitive for the robust component, however their use has been discouraged, as – for a given value of the mixture weight – it leads to full borrowing even in case of extremely large inconsistency between historical and concurrent data. This phenomenon is known in literature as Lindley's paradox. For this purpose, weakly informative robust components have been preferred and unit-information prior (UIP) has become a common choice, hence letting the weight of the mixture prior as the only parameter to be elicited based on the prior confidence of the sponsor in the external data. This choice poses some challenges, specifically *i*) the UIP's potential over-informativeness in trials with limited sample sizes, and *ii*) the inflation of the type I error (potentially up to 100%) under unequal allocation of patients to the control and experimental arms.

In this work we first prove that the posterior inference is driven by the specification of both the mixture weight and the robust variance, demonstrating in particular that (infinite many) different pairs of mixture weight and robust variance lead to (almost) the same posterior inference. Moreover, we prove that the joint selection of the mixture weight and variance of the robust component within a RMP framework effectively avoids incurring in Lindley's paradox, guaranteeing good borrowing properties even with arbitrarily large variances. We further demonstrate that employing large variance robust components mitigates type I error inflation in unbalanced trials (or even asymptotically eliminates it). The practical implications of all of these theoretical results will be demonstrated.

5: Bayesian Nonparametric Methods for Inferring Causal Effects of Longitudinal Treatments Amidst Missing Covariate Data

Liangyuan Hu

Rutgers University, United States of America

Background / Introduction Missing covariate data is a prevalent issue in longitudinal

studies, posing challenges for causal inference on longitudinal treatments. Imputation is a widely used solution, with most techniques relying on parametric models that explicitly define complex relationships among longitudinal responses, treatments, and covariates. However, incorrect specification of these parametric forms can lead to biases. While machine learning methods have gained traction for handling missing data, their development has predominantly focused on cross-sectional data, leaving longitudinal settings with repeated measures relatively underexplored.

Methods To address these limitations, we propose a flexible Bayesian nonparametric sequential imputation framework tailored for longitudinal data. We first develop a Bayesian ensemble-tree mixed-effects model BMTrees, and its variants, which leverage nonparametric priors to capture complex, non-linear relationships over time and handle non-normal random effects and errors. We then adapt BMTrees to the sequential imputation framework, effectively modeling relationships between observed and missing variables while incorporating a fitting-with-imputing strategy to enhance computational efficiency. This flexible imputation method can be seamlessly integrated with longitudinal causal inference approaches to enable coherent estimation of time-varying treatment effects.

Results Simulation studies demonstrate that BMTrees outperforms established ensemble-tree methods, including mixedBART and mixedRF, in both prediction and imputation tasks, particularly in challenging scenarios with non-normal data structures. Through a case study, we demonstrate the use of our sequential imputation method combined with noniterative conditional expectation estimator to evaluate the comparative effectiveness of antihypertensive treatment initiation thresholds for reducing long-term systolic blood pressure.

Conclusion We recommend using our proposed BMTrees method to impute longitudinal missing values, especially in scenarios where the dependence structure among longitudinal variables is nonlinear, and normality assumptions for model components are violated. This approach facilitates integrative causal analysis for evaluating time-varying treatment effects, coherently accounting for various sources of uncertainty.

Machine Learning 1

Tuesday, 2025-08-26 09:15 - 10:45, Biozentrum U1.101

Chair: Andreas Ziegler

1: Causal Machine Learning Methods for Dynamic and Static Treatment Strategies Deprescribing Medications in a Polypharmacy Population using Electronic Health Records.

Maurice M O'Connell¹, Michael Abaho², Aseel S Abuzour³, Matin Ahmed¹, Saiqa Ahmed⁵, Asra Aslam³, Danushka Bollegala², Iain Buchan², Harriet Cant¹, Andrew Clegg³, Mark Gabbay², Alan Griffiths⁵, Layik Hama³, Francine Jury¹, Gary Leeming², Emma Lo², Frances S Mair⁴, Simon Maskell², Erin McCloskey², Olusegun Popoola², Samuel Relton³, Roy A Ruddle³, Pieta Schofield², Eduard Shantsila², Tjeerd Van Staa¹, Lauren E Walker², Samantha A Wilson², Alan A Woodall², Rachael Wright², Matthew Sperrin¹

¹University of Manchester

²University of Liverpool

³University of Leeds

⁴University of Glasgow

⁵Public and Patient Partner

Introduction Despite recent advances, limited evaluation and guidance are available on the implementation of causal inference in the area of polypharmacy where high-dimensional confounding and medication interactions are present. The DynAIRx project (Artificial Intelligence for dynamic prescribing optimisation and care integration in multimorbidity) aims to develop statistical tools supporting GPs and pharmacists to find patients living with multimorbidity and polypharmacy who might be offered a better combination of medicines. We estimate treatment effects of discontinuing medications in a polypharmacy population.

Methods Within a polypharmacy cohort, we estimate the average causal effect of both time-fixed and time-varying treatment strategies comparing deprescribing versus continuing specific medications as advised by expert clinicians, e.g. individuals stopping either antiplatelets or anticoagulants, neither or both and the corresponding risk of strokes, bleeds and death. We underwent a detailed causal elicitation process with expert clinicians to draw causal diagrams, pre-specify dynamic treatment strategies and identify all important variables from electronic health records (EHRs).

We emulate target trials (using both sequential target trials and landmarking approaches) to estimate the average effect of treatment strategies using EHRs from the Clinical Practice Research (CPRD) Database. We target different causal estimands (e.g., total, direct, or separable effects) to allow for competing events from parametric pooled-over time logistic models using G-methods. Negative control outcomes are used to check robustness.

Causal machine learning is used to semi-parametrically include a larger combination of medications and interactions than included in our expert elicited causal diagram e.g., Targeted Maximum Likelihood Estimators, augmented inverse probability weighting with data adaptive approaches, cross-fitting, and super learner ensemble learning.

We plan to estimate individualised heterogeneous treatment effects (conditional average treatment effects over the smallest subgroups that can be supported by the data), e.g., R-, S-, T-, U-, X-, RS-, DR-learners evaluated with impact fraction rank-weighting metrics. When selecting patients for medication review, we can prioritise patients who are at highest risk of harms (benefits) from (changes in) treatment strategies.

Results We aim to present interim results from a large CPRD database consisting of millions of EHRs.

Discussion How do we give better advice in medication reviews to those with multimorbidity and polypharmacy, traditionally excluded from clinical trials? In clinical practice these prescribing decisions have been experienced a large number of times in EHRs. DynAIRx combines causal AI, guidelines and computing power linked to EHRs and where possible randomised controlled trials to estimate these complex causal effects.

2: Overview and Practical Recommendations on using Shapley Values for Identifying Predictive Biomarkers via CATE Modeling

David Svensson¹, Erik Hermansson¹, Konstantinos Sechidis², Nikos Nikolaou³, Ilya Lipkovich⁴

¹AstraZeneca, Sweden

²Novartis, Switzerland

³UCL, London

⁴Eli Lilly and Company, USA

Background / Introduction In recent years, two parallel research trends have emerged in machine learning: the modeling of Individual Treatment Effects, particularly the Condi-

tional Average Treatment Effect (CATE) using meta-learner techniques [1], and the field of Explainable Machine Learning (XML). While CATE modeling aims to identify causal effects from observational data, XML focuses on making complex models more interpretable, with Shapley Additive Explanations (SHAP) being a prominent technique [2]. Despite SHAP's popularity in supervised learning, its application in identifying predictive biomarkers through CATE models remains underexplored, especially in pharmaceutical precision medicine.

Methods We address the inherent challenges of applying SHAP in multi-stage CATE strategies by introducing an approach that is agnostic to the choice of CATE strategy, effectively reducing computational burdens in high-dimensional data. Our method involves a secondary modeling step after estimating individual treatment effects, regressing the estimated CATE against baseline covariates using a boosting model, from which SHAP importance is derived for each covariate.

Results Using our proposed method, we conduct simulation benchmarking to evaluate the ability to accurately identify biomarkers using SHAP values derived from various CATE meta-learners and Causal Forest [3]. This two-step approach provides a novel and unified way to evaluate different CATE models based on their ability to accurately identify predictive covariates through SHAP rankings. The results suggest that the architecture of the CATE model can greatly impact performance.

Conclusion Our study highlights key considerations when using SHAP values to explain models aimed at estimating causal quantities rather than traditional supervised learning. We investigate the operating characteristics of several popular CATE modeling choices using this new metric, providing insights into the impact of meta-learner schemes on covariate ranking accuracy. This research contributes to the understanding of predictive covariates underlying CATE estimates and offers a robust framework for future studies in causal inference and precision medicine.

References

- [1] Lipkovich I, Svensson D, Ratitch B, Dmitrienko A. Modern approaches for evaluating treatment effect heterogeneity from clinical trials and observational data. *Statistics In Medicine* 2024.
- [2] Lundberg S, Su-In L. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
- [3] Athey S, Imbens G. Recursive partitioning for heterogeneous causal effects. *PNAS*, 2016.

3: Signpost Testing to Navigate the High-Dimensional Parameter Space of the Linear Regression Model

Wessel van Wieringen^{1,2}

¹Dept. Epidemiology & Data Science, Amsterdam UMC

²Dept. Mathematics, Vrije Universiteit

Breast cancer knows many subtypes with different prevalences. While fundamentally different, the subtypes share commonalities. Such commonalities may benefit the statistical learning from data of the lesser prevalent subtypes, especially in high-dimensional situations. We evaluate the relevance of external quantitative information on the parameter of a linear regression model from high-dimensional data. The external information comes in the form of a parameter value available from a related knowledge domain or population, for instance from a more prevalent breast cancer subtype. The direction from a null to this externally provided parameter value serves as a signpost in the vast parameter space. We present a hypothesis test, the *signpost test*, to guide the search for the location of the parameter of the linear regression model in the high-dimensional setting. If the signpost test is significant, it is worthwhile to follow the signpost in the search of true parameter value. Our test statistic measures the relevance of the signpost's direction. We derive the test statistic's limiting distribution and provide approximations to other cases. The signpost's significance is assessed by comparing the signpost's direction to that of randomly rotations of this direction. We present a Bayesian interpretation of the signpost test and its connection to the global test. In simulation we investigate the signpost test's type I error and power, with particular interest in the effect of regularization and high-dimensionality in finite samples on these properties, and under misspecification of the alternative hypothesis. We employ the signpost test to illustrate how the learning of the regulatory mechanism of well-known cancer genes in a low prevalent breast cancer subtype benefits from external knowledge on this mechanism obtained from data of a more prevalent and related but fundamentally different subtype. The signpost test also finds use within the context of federated learning. It then serves as a means to evaluate the informativeness of an external parameter estimate, e.g. provided by a foreign institution, for the in-house model.

4: Leveraging Influence Functions for Statistical Inference in R

Klaus Holst

Novo Nordisk, Denmark

Influence functions (IFs), also known as influence curves or canonical gradients, are essential for characterizing regular and asymptotic linear estimators. They enable the direct calculation of properties such as asymptotic variance and facilitate the construction of new estimators through straightforward combinations and transformations. In this presentation, we will demonstrate how to work effectively with IFs in the statistical software R. Several examples will be provided to illustrate how to estimate and manipulate IFs, with specific applications in analyzing randomized clinical trials and multiple testing.

5: Optimal Testing for the Presence of Conditional Average Treatment Effects

Feng Liang¹, Kelly Van Lancker¹, Stijn Vansteelandt^{1,2}

¹Ghent University, Belgium

²London School of Hygiene and Tropical Medicine, UK

In May 2023, the U.S. Food and Drug Administration (FDA) issued industry guidance titled "*Adjustment for Covariates in Randomized Clinical Trials for Drugs and Biological Products*." This guidance advocates for the use of efficient estimators and tests for the average treatment effect, which adjust for baseline imbalances using flexible, data-adaptive methods while mitigating concerns about model misspecification bias. However, these approaches, though optimal for estimating the average treatment effect, do not fully leverage the information contained in baseline covariates, particularly in the presence of treatment effect heterogeneity, as the average treatment effect dilutes such heterogeneity.

To address this limitation, we develop optimal tests for the null hypothesis that the conditional average treatment effect is zero across all levels of measured baseline covariates. Our approach employs debiased machine learning and is inspired by the Projected Covariance Measure test, which we generalize to enhance its applicability. Through theoretical analysis and simulations, we compare our method with Maximum Likelihood Estimation and Augmented Inverse Probability Weighting, demonstrating that our test achieves higher statistical power while maintaining valid Type I error control. These advantages position our method as a competitive alternative for causal inference in both randomized clinical trials and observational studies, particularly in settings where treatment effect heterogeneity is anticipated.

Infectious Disease Modelling

Tuesday, 2025-08-26 09:15 - 10:45, ETH E27

Chair: Liesbeth C de Wreede

1: Estimating Transmission Parameters of Stochastic Epidemic Models using Survival Analysis Techniques

Hein Putter¹, Chengyuan Lu¹, Jacco Wallinga²

¹Leiden University Medical Center, Netherlands, The

²National Institute of Public Health and the Environment, Bilthoven; Netherlands, The

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 the present talk is to explain the link between stochastic epidemic models based on the susceptible-infectious-recovered (SIR) model, and methods from survival analysis. We develop new approaches to infer time-constant or time-varying transmission rates by building on the available work in the field of survival analysis. We first illustrate the basic ideas and terminology and notation in a highly idealized setting, with idealized data where all events in a population are observed, and an idealized stochastic epidemic model where all individuals are similar, except for their infection history, and where the transmission rate is constant. In this setting we can derive an explicit MLE estimator that we will use as a benchmark. We will show how time-varying transmission rates can be inferred and tests for constancy can be performed, and we will relax assumptions with respect to the underlying epidemic model, allowing for differences between groups of individuals and allowing for differences between individuals, additive terms such as import of infection, and multiplicative terms such as control measures that affect transmission rate. We will also suggest approaches to relax the assumptions made about the idealized data and allow for unobserved events, incompletely observed events, and binned data.

2: Parameter Estimation in Compartmental Epidemic Models with Heterogeneity in Susceptibility

Yuwen Ding¹, Jacco Wallinga^{1,2}, Hein Putter¹

¹Leiden University Medical Center, Leiden, The Netherlands

²National Institute for Public Health and the Environment, Bilthoven, The Netherlands

Introduction During the COVID-19 pandemic, compartmental epidemic models were crucial for forecasting epidemic dynamics and informing infection control policies. The susceptible-infectious-recovered (SIR) model provides a foundational framework for studying disease transmission but assumes a homogeneous population. To account for heterogeneity in susceptibility, we incorporate a frailty model by scaling the transmission parameter with individual random effects.

Methods This study focuses on estimating the transmission parameter and frailty variance. We consider three distributions from the power variance function (PVF) family — gamma, inverse Gaussian, and compound Poisson with probability mass at zero — in a completely observed epidemic scenario. Epidemic outbreaks are simulated in R, and the data are transformed into an individual-level representation. Using the Laplace transform of the PVF distribution, we derive the observed data likelihood function and obtain maximum likelihood estimates via the *optim* function. A profile Expectation-Maximization (EM) algorithm is also developed as an alternative approach.

Results For each frailty distribution, estimates for the transmission parameter and frailty variance are consistently close to the true values across various combinations of parameters and sample sizes, each evaluated over 1000 replications. The mean squared error remains small (<0.1), and the coverage of 95% confidence intervals closely aligns with the target level. The EM algorithm produces similarly accurate estimates but is computationally inefficient.

Conclusion Incorporating frailty distributions into the SIR model captures individual-level heterogeneity in susceptibility, providing a more nuanced representation of epidemic dynamics. The proposed estimation approach demonstrates both accuracy and efficiency. Future work will extend this method to handle more realistic data mechanisms, based on daily counts of new infections, and apply it to real-world epidemic scenarios.

3: Unobserved Intermediate Events in Multi-State Models in a Pandemic Setting

Ilaria Prosepe, Hein Putter, Mar Rodriguez-Girondo, Liesbeth C. de Wreede

Department of Biomedical Data Sciences, Leiden University Medical Center, the Netherlands

Background/Introduction Multi-state models are valuable for studying infectious diseases,

as they allow to study mortality accounting for various relevant events, interventions and covariates. In a pandemic setting, the most relevant intermediate events are infections and vaccinations. However, under-reporting of infections and (to a lesser extent) vaccinations is common. While the topic of under-reporting has received a fair amount of attention in infectious disease modeling, limited guidance exists on handling missing state transitions and transition times in multi-state models. In our work we aim to explore methodology to address this issue.

Methods We consider a semi-parametric (Cox-model based) illness-death model, with the intermediate state (illness) representing infection. The intermediate state is under-reported. In this model, our key estimands are the transition hazards, comprising baseline hazards and regression coefficients associated with relevant covariates, and the transition probabilities. Naïve estimation approaches that do not account for the under-reporting of infections underestimate the hazard of infection and overestimate the hazard of dying without illness. To overcome this, we explore how external knowledge can improve identifiability of each transition hazard by developing methods for incorporating it. Specifically, we investigate two approaches: recalibration of the naïve hazards by external data and multiple imputation of infection times with both data in the dataset and auxiliary external information on an aggregated level, such as hospitalization data.

Results We will report results from an extensive simulation study to evaluate how external information can be employed for the estimation of the transition hazards and transition probabilities in different settings. We will report point estimates, variance, empirical standard error, absolute bias and root mean square error.

Conclusions We present a framework of conditions under which the transition hazards of an illness-death model may be identified despite under-reporting of the intermediate state. We propose estimation methods and formulate an overview of pros and cons on how to proceed under different settings and under different types of external information. This helps to model the true burden of a pandemic.

4: Controlled Vaccine Efficacy using a Joint Model for Sparse Immunological Data and Time-to-Disease.

Grigorios Papageorgiou¹, Silvia Noirjean², Toufik Zahaf³, Andrea Callegaro³

¹GSK, Amsterdam, Netherlands, The

²GSK, Siena, Italy

³GSK, Wavre, Belgium

In vaccine development a key objective is to understand the immunological mechanisms that drive protection against the risk of disease. Typically, immune responses are collected at their expected peak level following the last vaccination dose planned. These peak measurements are subsequently used to establish correlates of protection (CoP) which means that the immunological biomarker can reliably predict vaccine efficacy (VE) and therefore act as a surrogate. There are several limitations and challenges that are typical in this setting. First, the immune response is expected to decay over time, and this might not be captured if only the peak measurements of the immunologic biomarker are used. Second, the sparsity in the collection of immune data over time poses an additional challenge in using approaches that exploit the whole immunological profile over time instead of the peak measurement.

To address these limitations and challenges we work under the joint modeling (JM) framework for longitudinal immunological data and time-to-disease data. This modeling approach enables us to use the whole immunological profile post vaccination and thus better understand the mechanisms that drive vaccine efficacy while improving its prediction. We view the sub-sampling of longitudinal responses as a missing data problem and show that under a missing at random (MAR) mechanism, inferences from the joint model are equivalent to analyzing the full cohort data if they were available. Furthermore, we leverage mediation analysis to define a causal effect called “controlled vaccine efficacy”, which captures the direct effect of the vaccine under different hypothetical immunological profiles. We show how this causal effect can be estimated using the joint model. Finally, we conduct a simulation study based on standard vaccine RCT settings to illustrate our approach, compare it with standard approaches and assess its performance under different scenarios and settings.

Our results suggest that the JM framework can overall improve the prediction of vaccine efficacy in terms of accuracy. The proposed JM controlled vaccine efficacy approach, enables the evaluation of longitudinal immune responses over time, rather than just peak levels, as a CoP.

5: Infectious Disease Estimands that are Insensitive to Interference

Mats Stensrud, Gellért Géza Perényi

EPFL, Switzerland

The treatment of one individual often affects outcomes other individuals. A canonical example occurs in infectious disease settings, where vaccinating one individual can reduce disease transmission and thereby affect the health outcomes of others. This type of interference implies that individuals cannot plausibly be perceived as independent and identically distributed

Abstracts of Contributed Talks

(iid). Extensive methodological research has recently been motivated by interference problems and the violation of conventional iid assumptions. However, despite growing interest in this topic, there remains controversy over whether and when existing methods capture causal effects of practical interest, such as in clinical medicine and public health.

In this talk, I will present causal methodology—motivated by infectious disease settings—for addressing interference. The central idea is to define estimands that are insensitive to the interference structure. This approach is not merely a workaround to avoid interference; rather, I will argue that the estimands have a clear interpretation and can guide decisions by doctors and patients. Specifically, these estimands can quantify vaccine waning and sieve effects, as illustrated by examples concerning COVID-19 and HIV.

Meta Science 1

Tuesday, 2025-08-26 09:15 - 10:45, ETH E23

Chair: Leonhard Held

1: Quantifying the Variation in Effects Due to Multiplicity of Analysis Strategies – a Conceptional Perspective

Susanne Strohmaier, S.Necdet Cervirme, Georg Heinze, Michael Kammer, Moritz Pamminger, Daniela Dunkler

Medical University of Vienna, Austria

When addressing a particular research question using observational data, many decisions must be made during the conceptualization of the statistical analysis plan (SAP). This *garden of forking paths* is a well-known problem leading to low replicability of research findings, as each decision can lead to different results, even if each decision on its own was scientifically justifiable.

In our ongoing project “Towards precise statistical analysis plans facilitating microdata analyses to advance health research” (TOPSTATS), we explore the variation in the relevant estimates in three case studies utilizing routinely collected Austrian data. These real-world applications originate from occupational epidemiology, nephrology, and pharmacoepidemiology, and rely on distinct time-to-event methodologies to address potentially causal research questions.

Several concepts have been proposed to tackle the consequences of the multiplicity of analysis strategies. For example, the social science literature promotes a *multi-model* approach to present a preferred model estimate in the context of results from other plausible models. This idea is similar to the concept of sensitivity analysis often used in epidemiological research, but focuses on the analysis models. The data science community is more concerned with *multiverse-style* methods aiming to neutrally present the results of ‘‘all’’ possible analysis decisions. Additionally, *multi-analyst* approaches examine the variation due to analytic choices deemed appropriate by independent analysts. We discuss the advantages and limitations of these existing concepts and suggest a *multi-analysis approach* that integrates elements from all these approaches, while maintaining the overarching goal of statistics: supporting decision-making in the face of uncertainty.

The core idea of TOPSTATS is to develop a consensus state-of-the-art (SOTA) SAP for

a relevant target estimand in each case study, as well as a meta-SAP comprising plausible alternative decisions at each step of an SAP together with international methodological experts. Data analysis then follows the SOTA-SAP and all sensible pathways through the meta-SAP. We suggest to present the result of our preferred (i.e., SOTA) analysis strategy in the context of a distribution of plausible estimates *while highlighting how decisions at different stages in the analysis path affect the estimate of interest.*

Motivated by our case studies, we present a pragmatic strategy how worthwhile paths through the landscape of meta-SAPs could be identified (a selection is necessary given the sheer number of possible paths) following ideas of a principled multiverse and discuss possible measures to empirically quantify the influence of particular decisions.

This work was supported through the ÖAW project DATA_2023-32_TopstatsMicrodata.

2: Uncertainty in Individual Predictions: Sample Size Versus Model and Modeler Choices

Toby Hackmann¹, Ben van Calster^{1,2,3}, Liesbeth C de Wreede¹, Ewout W Steyerberg^{1,4}

¹Department of Biomedical Data Sciences, LUMC, The Netherlands

²Department of Development and Regeneration, KU Leuven, Belgium

³EPI-Center, KU Leuven, Belgium

⁴Julius Center for Health Sciences and Primary Care, UMCU, The Netherlands

Background/Introduction Epistemic uncertainty in clinical prediction models is the uncertainty that can be estimated. It consists of sampling/approximation uncertainty and model-related uncertainty. The model-related source of uncertainty consists of uncertainty about the true model (model uncertainty) and about the knowledge, preferences and choices of the developer of the model (modeler uncertainty). We aim to quantify model and modeler uncertainty as separate from sampling uncertainty.

Methods As a case study, we developed prediction models for 30-day mortality after acute myocardial infarction as binary outcome based on data from the GUSTO-I trial. Models contained various subsets of potential predictors in different model classes. We quantify variability with a random-effects model with model-based predictions as outcome. Multiple model-based predictions are available for each patient arising from model and modeler choices. Sampling uncertainty was estimated through bootstrapping. Model categories (glm, RF, Neural Net) were modelled by the highest level random effect and within-category choices (hyperparameters, variable selection, regularization) led to a second-level random effect for

a total of 96 combinations. Combined with 100 bootstrap replications, 9600 predictions were made per patient. Model choices were based on a scoping review of common modeling choices. Within-category choices were optimized using common performance metrics. Sample sizes ranged between 400, 2,000, and 10,000 patients with 7% experiencing the event of interest.

Results The variance between predictions for an individual patient based on different model categories and/or within-category choices was substantial. Preliminary results showed that within-category standard deviation for predictions based on logistic regression models was 0.005, compared to standard errors due to sampling of 0.024 for 1,000 patients and 0.011 for 5,000 patients.

Conclusions Model and modeler uncertainty are major additional sources of uncertainty over sampling uncertainty, which may be uncovered by bootstrapping. Following guidance on good practices could reduce the variability between model choices and make predictions for individual patients more stable between different modelers.

Acknowledgements We would like to thank Napsugar Forró, MSc for her help in developing the mixed-model approach for the estimation of between-model uncertainty.

3: Quantifying Reproducibility - Results from a Scoping Review on Reproducibility Metrics and Simulation Studies into their Real-World Applicability

Rachel Heyard¹, Samuel Pawel¹, Joris Frese², Bernhard Voelkl³, Hanno Würbel³, Sarah K McCann⁴, Kimberley E Wever⁵, Helena Hartmann⁶, Louise Townsin⁷, Stephanie Zellers⁸, Leonhard Held¹

¹University of Zurich, Switzerland

²European University Institute, Florence, Italy

³University of Bern, Bern, Switzerland

⁴QUEST Center, BIH, Berlin, Germany

⁵Radboud University Medical Center, Nijmegen, the Netherlands

⁶University Hospital Essen, Essen, Germany

⁷Torrens University Australia, Australia

⁸University of Helsinki, Helsinki, Finland

Background Replication studies and large-scale replication projects aiming to quantify different aspects of reproducibility (such as the Reproducibility Project: Cancer Biology) are increasingly common. The iRISE (improving Reproducibility In SciencE) consortium defines

reproducibility as “the extent to which the results of a study agree with those of replication studies”. Currently, no standardized approach to measuring reproducibility exists and a diverse set of metrics is in use. Further, little is known about the applicability and performance of reproducibility metrics under various conditions and in real-world contexts.

Methods To identify reproducibility metrics, we conducted a scoping review of large-scale replication projects that used metrics and methodological papers that suggested or discussed them. A list of 49 large-scale projects was compiled by the research team, and 97 methodological papers were identified through a search in Scopus, MedLine, PsycINFO and EconLit. To study the metrics’ applicability in various real-world contexts, simulation studies and real-world data analyses were performed. We were specifically interested in the applicability of metrics, initially developed to assess whether a direct replication study was successful, in the context of the translation of results from preclinical animal studies to human trials. To ensure the practical relevance of this translation simulation study, we used real-world data to select simulation parameters.

Results We identified 50 reproducibility metrics and characterized them based on their type (e.g. formulas and/or statistical models, graphical representations, algorithms), input required, and appropriate application scenarios. We found that each metric addresses a distinct research question. Preliminary results from our simulation study indicate that specific metrics are more useful in some contexts compared to others, showing the importance of the choice of the metric for the validity of (large-scale) replication projects.

Conclusion To support future replication teams and meta-researchers, we provide a comprehensive and interactive “live” table to guide the selection of the most appropriate metrics aligned with goals of the study or large-scale project. We present assumptions and limitations of some commonly used metrics and give tangible recommendations for their application in various contexts, including the translation of results from preclinical animal studies to human trials.

4: Data Quality, a Blind Spot in Study and Reporting Guidelines

Carsten Oliver Schmidt

University Medicine Greifswald, Germany

Background

Ensuring that data is ‘fit for purpose’ is fundamental to credible, replicable and reproducible scientific research. Numerous works discuss concepts and tools related to data quality. How-

ever, does this translate to explicit requirements on transparent data quality reporting in reporting guidelines, appraisal tools, and journal author instructions? This work critically examines how such guidelines address data quality.

Methods

A comprehensive review of key reporting guidelines, a review of appraisal tools², and journal author instructions was conducted to assess the extent to which data quality is explicitly considered. The analysis included for example major guidelines such as CONSORT, SPIRIT, STROBE, PRISMA, STARD, TRIPOD, and TRIPOD-AI, as well as journal submission guidelines from high-impact medical journals. The review focused on explicit mentions of data quality, as well as the coverage of its key constituting elements, including data integrity, completeness, correctness, and variability.

Results

Findings indicate a striking lack of emphasis on data quality within existing guidance documents. Among the major reporting guidelines, only limited and vague references to 'data quality' or related processes exist, often without concrete definitions or structured reporting requirements. There was only one single mention of 'data quality' across hundreds of evaluation criteria in 49 appraisal tools. Key data quality components such as measurement error, misclassification, and heterogeneity are inconsistently addressed, with most guidelines omitting them entirely. A review of journal author instructions revealed similarly limited references to data quality, with most guidance focusing on study design and statistical methods rather than providing empirical evidence on the quality of the underlying data.

Conclusion

Transparency regarding data quality should be the norm, not the exception. However, the absence of any systematic coverage of data quality-related aspects represents a major shortcoming in current guidance practice. Consequently, the reverse is the case: transparency remains the exception rather than the norm¹. To address this fundamental flaw, structured and transparent data quality reporting should be a key component of all relevant guidelines and should be routinely enforced by journals and funding agencies.

1. Huebner et al., & Topic Group "Initial Data Analysis" of the, S. I. (2020). Hidden analyses: a review of reporting practice and recommendations for more transparent reporting of initial data analyses. *BMC Medical Research Methodology*, 20(1), 61. doi:10.1186/s12874-020-00942-y
2. Jiu et al. Tools for assessing quality of studies investigating health interventions using real-world data: a literature review and content analysis. *BMJ Open* 2024; 14(2): e075173.

5: To Adjust or Not: it is not the Tests Performed that Count but How They are Reported and Interpreted

Sabine Hoffmann, Simon Lemster, Juliane Wilcke, Anne-Laure Boulesteix

LMU Munich, Germany

Most original articles published in the medical literature report the results of multiple statistical tests. In a few simple cases, there is general agreement on whether one should adjust for multiple testing. For many cases encountered in practice, however, this is less clear and the recommendations in the literature are contradictory, along different dimensions, or otherwise confusing. This lack of clear guidance may hinder or impair the conduct of analyses, encourage questionable research practices, ultimately jeopardizing the credibility of medical research. In this project, we present a unifying criterion that supports researchers in their decision to adjust or not for multiple testing and if yes over which set of hypotheses. We relate the criterion to previous rules proposed in the literature and illustrate its use in two complex multiple testing situations. In addition, we show that the criterion also addresses the multiple testing situation resulting from the multiplicity of possible analysis strategies for a given research question and dataset, which - if not handled properly - opens the door to fishing for significance and false positive findings.

Design and Analysis of Trials in Rare Diseases

Tuesday, 2025-08-26 09:15 - 10:45, ETH E21

Chair: Martin Posch

1: Blinded Sample Size Re-Estimation Accounting for Estimation Error with Small Internal Pilot Studies

Hirotada Maeda¹, Satoshi Hattori¹, Tim Friede²

¹Graduate School of Medicine Division of Medicine, Osaka University

²Department of Medical Statistics, University Medical Center Göttingen

In randomized controlled trials, it is important to set the target sample size accurately at the design stage for the study to be conclusive. However, the sample size formula often requires specification of some parameters other than the treatment effect, which are often referred to as nuisance parameters. Their misspecification can lead to studies with insufficient power. For example, in comparing two normal populations, specification of the standard deviation is influential to the power of the final analysis. Blinded sample-size re-estimation is an approach to avoid inaccurate sample size calculation. Kieser and Friede (2003) proposed to use the one-sample variance, which can be estimated in a blind review without knowledge of the treatment allocation. We point out that their method is regarded as a worst-case evaluation with the largest variance subject to information under blinded treatment allocation and then can likely avoid underpowered studies. However, as they reported, with blinded reviews of small sample size, it still may lead underpowered studies. We propose a refined method accounting for estimation error in blind reviews using confidence intervals of the one-sample variance. We developed a method to select an appropriate confidence level so that the re-estimated sample size attains the target power. The idea is related to the sample size calculation method with pilot studies by Kieser and Wassmer (1996). The required confidence level can be prespecified in the protocol and coupled with the blinded one-sample variance estimate, one can determine the sample size for the final analysis of the target power to detect the pre-specified treatment effect, maintaining study integrity. We conducted numerical studies to evaluate the performance of the proposed method and concluded that our method worked well as designed and outperformed existing methods.

2: Randomization in Clinical Trials with Small Sample Sizes using Group Sequential Designs

Daniel Bodden¹, Ralf-Dieter Hilgers¹, Franz König²

¹Institute of Medical Statistics, RWTH Aachen University, Aachen

²Institute of Medical Statistics, Center for Medical Data Science, Medical University of Vienna

Background Group sequential designs, which allow early stopping for efficacy or futility, may benefit from balanced sample sizes at interim and final analyses. This requirement for balance limits the choice of admissible randomization procedures. We investigate if the choice of randomization procedure, whether balanced or not, impacts the type I error probability and power in trials with group sequential designs.

Methods We investigate the impact of randomization procedures on the type I error probability and power of trials with Pocock, O'Brien-Fleming, Lan-DeMets and inverse normal combination test designs.

Results Simulation results demonstrate that deficiencies in the implementation of randomization can inflate type I error rates. Some combinations of group sequential designs and randomization procedures cause a loss of power, for example, when using inverse normal combination tests.

Conclusion We propose a framework for selecting the most suitable combinations of group sequential design and randomization procedure. When the planned balanced allocation ratio in (interim) analyses cannot be ensured, the Lan-DeMets approach is preferable for small sample trials due to its robustness to deviations between the planned and observed allocation ratio. The inverse normal combination test, while useful in trials with limited prior information, should be used cautiously with permuted block randomization that maintains the planned allocation ratio to avoid power loss.

3: Adjusting for Allocation Bias in Stratified Clinical Trials with Multi-Component Endpoints

Stefanie Schoenen¹, Ralf-Dieter Hilgers¹, Nicole Heussen²

¹RWTH Aachen, Germany

²Sigmund Freud Private University, Vienna, Austria

Background The disease heterogeneity and geographic dispersions of patients in rare diseases often necessitates a multi-centre design and the use of multi-component endpoints, which combine multiple outcome measures into a single score. A common issue in these trials is allocation bias, as trials in rare diseases are frequently unblinded or single-blinded. Allocation bias occurs when future allocations can be predicted based on previous ones, potentially leading to the preferential assignment of patients with specific characteristics to either the treatment or control group. The ICH E9 guideline recommends assessing the potential contributions of bias to inference. Therefore, our research aims to develop a bias-adjusted analysis strategy for stratified clinical trials with multi-component endpoints.

Methods To model biased patient responses, we derived an allocation biasing policy based on the convergence strategy of Blackwell and Hodges [1], which assumes that the next patient will be allocated to the group with fewer prior assignments. Using this policy, we formulated a bias-adjusted analysis strategy for a stratified version of the Wei-Lachin test, which is a combination of Fleiss's stratified test and the Wei-Lachin test [2,3].

Through simulations, we assess the impact of allocation bias on the type I error rate of the stratified Wei-Lachin test, both with and without bias adjustment, and evaluate how statistical power is affected when accounting for allocation bias.

Results Allocation bias increases the type I error rate of the stratified Wei-Lachin test, potentially exceeding the 5% significance level. Therefore, if allocation bias is a concern, a bias-adjusted analysis should be conducted as a sensitivity analysis to ensure valid results. The bias-adjusted stratified Wei-Lachin test maintains the 5% significance level while preserving approximately 80% power under both unbiased and biased conditions. In contrast, the unadjusted test shows an inflated power exceeding 80% in the presence of bias, leading to an overestimation of the true treatment effect.

Conclusion Conducting a bias-adjusted test as sensitivity analysis improves the validity of trial results. The proposed methodology enhances the robustness of rare disease clinical trials, ensuring more reliable and accurate conclusions.

[1] Blackwell D, Hodges JL. Design for the Control of Selection Bias. *The Annals of Mathematical Statistics*. 1957;28(2):449–60.

[2] Fleiss JL. Analysis of data from multiclinic trials. *Controlled Clinical Trials*. 1986;7(4):267–275.

[3] Wei LJ, Lachin JM. Two-Sample Asymptotically Distribution-Free Tests for Incomplete Multivariate Observations. *Journal of the American Statistical Association*. 1984;79(387):653–661.

4: Methodological Insights from the EPISTOP Trial for Designing and Analysing Clinical Trials in Rare Diseases

Stephanie Wied, Ralf-Dieter Hilgers

RWTH Aachen University, Germany

Background The most suitable method for assessing the impact of an intervention in clinical research is to conduct a randomised controlled trial (RCT). However, implementing an RCT can be challenging, especially in small population groups. These challenges can arise during the planning phase of a clinical trial and may occur later, when potential solutions may no longer be feasible. The EPISTOP trial aimed to compare outcomes in infants with tuberous sclerosis (TSC) who received vigabatrin preventively before seizures with those who were treated conventionally after seizure onset [1]. The study was designed as a prospective, multicentre, randomised clinical trial. However, ethics committees at four centres did not approve this RCT design, resulting in an open-label trial (OLT) in these four centres.

Methods We investigate whether randomisation introduced any bias in the EPISTOP trial and how to address the presence of different types of data (RCT and OLT data) within the context of clinical trials. To support and strengthen the published results, we re-analyse the data from the EPISTOP trial using a bias-corrected analysis [2]. The statistical model includes a term representing the effect of selection bias as a factor influencing the corresponding endpoint. As a result, the treatment effect estimates for the primary endpoint of time to first seizure, as well as secondary endpoints are adjusted for the impact of bias.

Results The bias-corrected analyses for the primary endpoint indicate quite similar estimated hazard ratios and associated confidence intervals for original and bias-corrected analysis (original: HR 2.91, 95%CI [1.11 to 7.67], p-value 0.0306; bias-corrected: HR 2.89, 95%CI [1.10 to 7.58], p-value 0.0316). This consistency was also observed in the secondary endpoints. Therefore, the statistical reanalysis of the raw study data supports the published results and does not demonstrate additional bias related to randomisation.

Conclusion In summary, it becomes clear that the prevention and quantification of bias should be taken into account in future clinical studies to ensure reliable study results.

References [1] Kotulska, K. et al. (2020), Prevention of Epilepsy in Infants with Tuberous Sclerosis Complex in the EPISTOP Trial. Ann Neurol, 89: 304-314. <https://doi.org/10.1002/ana.25956>

[2] Wied, S. et al. (2024) Methodological insights from the EPISTOP trial to designing clinical trials in rare diseases - A secondary analysis of a randomized clinical trial. PLOS ONE 19(12). <https://doi.org/10.1371/journal.pone.0312936>

5: Modified Crossover Trials to Improve Feasibility of Evaluating Multiple Treatments for Rare Relapsing-Remitting Conditions

James Wason

Newcastle University, United Kingdom

Background It is challenging to conduct well-powered clinical trials for rare diseases due to the limited number of patients available to recruit. Trial designs that are statistically efficient and appealing to potential trial participants make a big difference to how feasible the trial is to conduct.

For chronic relapsing-remitting conditions, crossover trials are well-established for treating participants with multiple interventions, in sequence. They are highly statistically efficient, however may be off-putting to participants as they involve stopping a treatment at a specified point, even if it is providing benefit.

This presentation discusses a modified crossover trial design, developed for the BIOVAS trial. BIOVAS assessed the effect of three biologic therapies vs placebo for patients with non-ANCA associated vasculitis. The design used a time-to-event outcome representing occurrence of disease flare (recurrence of symptoms), with participants moving on to the next treatment in the sequence after flare occurs. In this way, participants remain on a treatment whilst they are benefitting.

Methods Using a simulation study, the statistical properties of the trial design are shown assuming a mixed-effects Cox regression model is used to analyse the trial. Considerations on how blinding can be implemented are provided. The development of two newer, in VEXAS syndrome and Juvenile Scleroderma, using a similar design will be highlighted.

Results Simulation studies showed no evidence of type I error rate inflation or non-negligible statistical bias in realistic situations. Careful consideration of blinding is necessary to ensure participants do not become unblinded during the sequence.

Conclusion This modified crossover design improves patient acceptability by allowing continued benefit from treatment while maintaining high statistical efficiency.

Survival and Recurrent Events in Clinical Trials

Tuesday, 2025-08-26 11:30 - 13:00, Biozentrum U1.131

Chair: Theis Lange

1: Reevaluating Recurrent Events in Heart Failure Trials: Patterns, Prognostic Implications, and Analytical Improvements

Audinga-Dea Hazewinkel¹, John Gregson¹, Stuart J Pocock¹, John McMurray², Scott D Solomon³, Brian Claggett³, Milton Packer^{4,5}, Stefan D Anker^{6,7,8}, João P Ferreira^{9,10,11}, Kieran Docherty², Alasdair Henderson², Pardeep Jhund², Ulrica Wilderäng¹², David Wright¹³

¹Department of Medical Statistics, London School of Hygiene & Tropical Medicine, London, UK

²BHF Cardiovascular Research Centre, University of Glasgow, London, UK

³The Cardiovascular Division, Brigham and Women's Hospital, Boston, USA

⁴Baylor University Medical Center, Dallas TX, USA

⁵Imperial College, London, UK

⁶Department of Cardiology (CVK) of German Heart Center Charité, Universitätsmedizin, Berlin, Germany

⁷Berlin Institute of Health Center for Regenerative Therapies (BCRT), Berlin, Germany

⁸German Centre for Cardiovascular Research (DZHK) partner site Berlin, Charité Universitätsmedizin, Berlin, Germany

⁹Department of Surgery and Physiology, Cardiovascular Research and Development Center (UnIC@RISE), Faculty of Medicine of the University of Porto, Porto, Portugal

¹⁰Heart Failure Clinic, Internal Medicine Department, Unidade de Saude de Gaia-Espinho, Portugal

¹¹Université de Lorraine, Inserm, Centre d'Investigations Cliniques, - Plurithématique 14-33, Inserm U1116, CHRU Nancy, F-CRIN INI-CRCT, France

¹²R&D BioPharmaceuticals, Late-Stage Development, Cardiovascular, Renal and Metabolism (CVRM), AstraZeneca, Gothenburg, Sweden

¹³Head of Statistical Innovation, Respiratory and Immunology Biometrics and Statistical Innovation, Biopharmaceuticals R&D, AstraZeneca, Cambridge, UK

Background Repeat events analyses are sometimes used as the primary outcome in randomized trials in heart failure, often due to a perception that including more events will increase statistical power. Such analyses frequently use negative binomial or LWYY models – extensions of Poisson and Cox regression. These models allow for variation in underlying

disease risk between patients, but ignore the time-dependency of events within a patient. We aimed to assess (i) the time-clustering of repeat outcomes over time; (ii) whether using repeat events analyses improves statistical power

Methods We use data from four large trials in chronic heart failure (EMPEROR-Reduced, EMPEROR-Preserved, DAPA-HF and DELIVER), each with a primary outcome of heart failure hospitalisation (HFH) or cardiovascular death. We explore the time-clustering of outcome events and develop a method for quantifying it. We compare the use of time-to-first event versus recurrent events for estimating treatment benefit. To do so we use two non-parametric methods: the win ratio method and the area under the curve (AUC) method – an extension of restricted mean survival time for repeat events.

Results With 20,725 patients in all four trials, there were 1,844 cardiovascular deaths, and 3,913 HFHs in 2,494 patients. The distribution of HFHs was highly skewed, with a few patients having many events. The mean time between consecutive HFHs within a patient was 20%-24% shorter than if they occurred randomly over time. Following an HFH, subsequent rehospitalization for heart failure or cardiovascular death were markedly more common, particularly in the first 3 months post-discharge (adjusted hazard ratios: 5.5 to 10.5 and 6.3 to 8.8, respectively). While decreasing over time, the risk of subsequent rehospitalization or cardiovascular death remains significantly elevated, even after 1 year post-discharge. In analyses using the win ratio and AUC, using repeat event analyses rather than time-to-first event led to a less statistically significant estimate of treatment effect in 3 out of 4 trials. Similar findings occurred using negative binomial and LWYY models, though the value of these approaches can be questioned given they do not take account of time-clustering.

Discussion The perception that using repeat events rather than time-to-first event gains statistical power in heart failure trials appears to be misplaced. This appears due to the highly skew distribution of repeat events, and the strong time-clustering of events within patients. We looked at drug treatment trials in chronic heart failure; findings may be different for other trial types.

2: Balancing Events, not Patients, Maximizes Power of the Logrank Test: And Other Insights on Unequal Randomization in Survival Trials

Godwin Yung¹, Kaspar Rufibach², Marcel Wolbers¹, Ray Lin¹, Yi Liu³

¹F. Hoffmann-La Roche

²Merck Group

³Nektar Therapeutics

We revisit the question of what randomization ratio (RR) maximizes power of the logrank test in event-driven survival trials under proportional hazards (PH). By comparing three approximations of the logrank test (Schoenfeld, Freedman, Rubinstei) to empirical simulations, we find that the RR that maximizes power is the RR that balances number of events across treatment arms at the end of the trial. This contradicts the common misconception implied by Schoenfeld's approximation that 1:1 randomization maximizes power. Besides power, we consider other factors that might influence the choice of RR (accrual, trial duration, sample size, etc.). We perform simulations to better understand how unequal randomization might impact these factors in practice. Altogether, we derive 5 insights to guide statisticians in the design of survival trials considering unequal randomization.

3: An Alternative to Classical Intention-to-Treat Analysis for Comparing a Time-to-Event Endpoint in Precision Oncology Trials

Marilena Müller

DKFZ Heidelberg, Germany

Background / Introduction We consider a two-arm randomized clinical trial in precision oncology with time-to-event endpoint. The control arm consists of standard of care (SOC) whereas patients in the treatment arm are offered personalized treatment, when available. Patients in the personalized treatment arm, for which no personalized treatment is available or who do not consent to treatment also receive SOC. Intention-to-treat analysis hence involves comparing the outcomes of a group of patients receiving either personalized treatment or SOC to patients receiving exclusively SOC. This does not lead to an unbiased estimator of the treatment effect for those eligible for personalized treatment in the classical intention-to-treat approach.

Methods We investigate the performance of intention-to-treat and per-protocol analyses and develop more appropriate alternative analysis schemes. For this purpose, the patients are divided into groups based on whether they receive their intended treatment or not. An extension of the Cox proportional hazards model is proposed for estimating the conditional intensities in each group simultaneously via Maximum Likelihood estimation on the partial likelihoods. Counting process theory as well as martingale theory is used to develop suitable test statistics for various settings of interest. Both groups can be evaluated distinctly, thus

enabling comparison between groups. This includes the investigation of a possible selection effect via the groups' respective regression coefficients.

Results A regression model is proposed that allows for modelling the differences between complying patients and non-complying patients, which enables the evaluation of the selection effect in the case of asymmetric trial arms. An in-depth simulation study and a real data example complement the theoretical results.

Conclusion A novel more rigorous model for the analysis of the treatment effect in the presence of mixtures or asymmetric trials is proposed. Guidelines are provided to identify scenarios where this model is necessary or appropriate, and when a classical intention-to-treat analysis remains preferable.

4: Proposing a New Method to Estimate the Survival of the Intention-to-Treat Population in Trials with Two-Stage-Randomization-Design

Dan Huang^{1,2}, Eva Hoster², Stefan Englert¹, Martin Dreyling³, Ulrich Mansmann²

¹Janssen Research & Development, Janssen-Cilag GmbH, a Johnson & Johnson company, Neuss, Germany

²Institute for Medical Information Processing, Biometry and Epidemiology (IBE), LMU Munich, Munich, Germany

³Department of Internal Medicine III, LMU University Hospital Munich, Munich, Germany

Background Two-Stage-Randomization-Design (TSRD) trials are common in clinical research, where patients firstly randomized between two induction regimens followed by another randomization to different maintenance therapies conditional on response to induction. In cancer trials using TSRD, patients seek to understand the failure-free survival (FFS) and overall survival (OS) for a specific combination of induction and maintenance therapies. However, the primary analysis usually focuses on survival estimates of the maintenance part limited to patients randomized in maintenance phase, neglecting the efficacy of adaptive treatment combinations of induction and maintenance according to the intention-to-treat (ITT) principle. Inverse-Propensity -Weighting (IPW) has been proposed to estimate the survival including non-randomized patients. However, it relies heavily on the accurate selection of baseline characteristics and assumes no confounding factors are present. We propose a novel and a simpler method for estimating the intent-to-treat effect: retrospective randomization.

Methods The principle of retrospective randomization is to mimic a study design of prospec-

tive, up-front randomization. In our approach, the non-randomized patients undergo retrospective randomization between maintenance arms during the statistical analysis steps, using the same stratification factors employed in the patients prospectively randomized. This process of retrospective randomization is repeated multiple times. The FFS and associated hazard ratios are then estimated by combining the results obtained from each iteration, including patients prospectively and retrospectively randomized for maintenance treatment.

Clinical data with TSRD were simulated to contain patients' demographics, response status and time to response to the induction therapy and FFS from the maintenance phase randomization. The disease response rate to the induction therapy, randomization proportion out of all responders and the outcome by remission status varied across scenarios. For comparison purpose, the IPW method was also applied for the FFS estimation. The performance of the proposed retrospective randomization method was evaluated using bias and root mean squared error (RMSE) compared to the true estimates of FFS. Analysis was also performed in a double randomized Mantle Cell Lymphoma (MCL) elderly trial of the European MCL Network.

Results and conclusion Across different simulation scenarios and in the case study, the retrospective randomization method delivers survival estimates close to the actual values of the intention-to-treat population and a more meaningful estimate than the estimates provided in primary maintenance analysis. Our method achieves comparable or even better performance in some scenarios than the IPW method in terms of bias and RMSE, while offering advantages in implementation simplicity and reduced reliance on assumption.

5: RCT for Recurrent Events

Thomas Scheike

University of Copenhagen, Denmark

We consider how to compare treatments based on a randomized clinical trial (RCT) when the outcome of interest is the number of recurrent events. The interest from the medical perspective is to find the treatment that leads to the lowest expected number of recurrent events. This question can often be addressed by using Andersen-Gill type models with robust standard errors. The efficiency can be improved by using auxiliary covariate information that is often available. We show how this can be accomplished by extending the augmentation approach of Lu & Tsiatis (2008a), and that this ensures robustness to misspecifications when testing for no-treatment effects. The efficiency gain obtained from auxiliary covariates is closely related to the use of covariate adaptive randomization techniques, and we also

Abstracts of Contributed Talks

point out how to compute standard errors when the RCT is based on such techniques (Ye & Shao, 2018; Bugni et al., 2018). Further, we demonstrate that the efficiency gain can be large and can be obtained without relying on any modeling assumptions. The techniques are shown to work in simulations and illustrated in practical use by the RCT that motivated the work. In particular we develop an RCT augmented baseline estimator.

Alternative Estimands in Causal Inference

Tuesday, 2025-08-26 11:30 - 13:00, Biozentrum U1.141

Chair: Elise Dumas

1: Assumption-Lean Modeling for Patient-Centered Decision-Making

Stijn Vansteelandt, Georgi Baklicharov

Ghent University, Belgium

The use of odds ratios and hazard ratios as primary effect measures in randomized trials and observational studies has faced significant criticism due to challenges in interpretation and communication. Recent advancements have shifted focus toward model-free estimands for marginal treatment effects, employing data-adaptive statistical modeling and machine learning to address imbalances in randomized trials or confounding in observational studies. However, these methods often lack direct applicability to patient-centered decision-making. While conventional treatment effect measures, such as odds ratios, can map conditional counterfactual mean outcomes (given baseline covariates) from untreated to treated scenarios, this involves reliance on parametric structures, making them susceptible to bias from model misspecification.

We address this challenge by introducing a novel debiased machine learning strategy that minimizes the squared bias induced by such mappings. Our approach ensures minimal bias by construction, leverages flexible data-adaptive methods to estimate conditional outcome means, accommodates model uncertainty (e.g., variable selection uncertainty), and provides insights into subject-specific treatment effects, enabling transparent communication with patients and clinicians through tailored visualizations.

Our proposal drastically advances previously proposed assumption-lean modeling strategies by maintaining interpretability under model misspecification, delivering more efficient estimators, and offering insight into treatment effect heterogeneity. We demonstrate the advantages of our method through simulation studies and an analysis of a recent diabetes trial.

2: Causal Inference Targeting a Concentration Index for Studies of Health Inequalities

Mohammad Ghasempour, Xavier de Luna, Per Gustafsson

Umeå University, Sweden

A concentration index, a standardised covariance between a health variable and relative income ranks, is often used to quantify income-related health inequalities. There is a lack of formal approach to study the effect of an exposure, e.g., education, on such measures of inequality. In this paper we contribute by filling this gap and developing the necessary theory and method. Thus, we define a counterfactual concentration index for different levels of an exposure. We give conditions for the identification of this complex estimand, and then deduce its efficient influence function. This allows us to propose estimators, which are regular asymptotic linear under certain conditions. In particular, we show that these estimators are \sqrt{n} -consistent and asymptotically normal, as well as locally efficient. The implementation of the estimators is based on the fit of several nuisance functions. The estimators proposed have rate robustness properties allowing for convergence rates slower than \sqrt{n} -rate for some of the nuisance function fits. The relevance of the asymptotic results for finite samples is studied with simulation experiments. We also present a case study of the effect of education on income-related health inequalities for a Swedish cohort.

3: Rethinking the Win Ratio: A Causal Framework for Hierarchical Outcome Analysis

Julie Josse, Mathieu Even

Inria, France

Quantifying causal effects in the presence of complex and multivariate outcomes is a key challenge to evaluate treatment effects. For hierarchical multivariate outcomes, the FDA recommends the Win Ratio and Generalized Pairwise Comparisons approaches. However, as far as we know, these empirical methods lack causal or statistical foundations to justify their broader use in recent studies. To address this gap, we establish causal foundations for hierarchical comparison methods. We define related causal effect measures, and highlight that depending on the methodology used to compute Win Ratios or Net Benefits of treatments, the causal estimand targeted can be different, as proved by our consistency results. Quite dramatically, it appears that the causal estimand related to the historical estimation approach

can yield reversed and incorrect treatment recommendations in heterogeneous populations, as we illustrate through striking examples. In order to compensate for this fallacy, we introduce a novel, individual-level yet identifiable causal effect measure that better approximates the ideal, non-identifiable individual-level estimand. We prove that computing Win Ratio or Net Benefits using a Nearest Neighbor pairing approach between treated and controlled patients, an approach that can be seen as an extreme form of stratification, leads to estimating this new causal estimand measure. We extend our methods to observational settings via propensity weighting, distributional regression to address the curse of dimensionality, and a doubly robust framework. We prove the consistency of our methods, and the double robustness of our augmented estimator. These methods are straightforward to implement, making them accessible to practitioners.

4: Assessing Individual-Level Uncertainty of Causal Predictions Through the Causal Effective Sample Size

Doranne Thomassen¹, Daniala Weir², Marleen Kunneman^{3,4}, Nan van Geloven¹

¹Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands

²Division of Pharmacoepidemiology and Clinical Pharmacology, Department of Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands

³Department of Public Health and Primary Care, Leiden University Medical Center, Leiden, The Netherlands

⁴Knowledge and Evaluation Research Unit, Mayo Clinic Rochester, Rochester, MN, USA

Background As prediction algorithms are increasingly used to support individual decision making in healthcare, the fairness, trustworthiness and transparency of such algorithms are paramount. To prevent biases of regular prediction algorithms developed on observational data, causal prediction (also referred to as counterfactual prediction or prediction under interventions) has been proposed as a basis for decision support. Causal prediction relies on untestable assumptions about the data underlying the algorithm. In view of algorithm trustworthiness, the plausibility of these assumptions should be evaluated as well as the statistical uncertainty associated with the algorithm. Crucially, both can vary widely between individual patients. We aimed to develop a metric that helps to evaluate and communicate the trustworthiness of a causal prediction algorithm on an individual patient level.

Methods For regular prediction algorithms, the effective sample size for an individual can be interpreted as the number of similar individuals (i.e., with similar predictor values) that their prediction is effectively based on. To translate the concept of effective sample size

to a causal prediction setting, we explored the connections between the effective sample size and the causal inference assumptions of exchangeability and positivity. Furthermore, we developed methods to estimate effective sample sizes for causal predictions. These methods were applied to a clinical dataset, leading to a prototype of how causal effective sample size might be presented to end-users as part of a decision aid used in diabetes care.

Results In contrast to the regular effective sample size, the causal version may consist of multiple effective sample sizes for the same individual, i.e. one for each (hypothetical) intervention option considered. In addition, the causal effective sample size should extend the definition of ‘similar individuals’ to cover adjustment factors (confounders) that are not part of the predictor set. As such, the causal effective sample size can be used to consider the plausibility of the positivity assumption for an individual. In the clinical data, we observed large between-individual differences in causal effective sample sizes.

Conclusion While causal assumptions are traditionally only evaluated globally, we have shown how they can also be evaluated for individuals. This improves the transparency of causal prediction algorithms and allows end-users to determine on an individual basis whether they trust and want to use the predictions from the algorithm in their decision-making.

Funding This project has received funding from the Dutch Organisation for Scientific Research (NWO) under Grant ID (DOI) 10.61686/DFECP93059.

5: Estimating Win Ratio for Prioritized Composite Outcomes in the Presence of Noncompliance

Md. Muhitul Alam¹, Mahbub A.H.M. Latif¹, M. Iftakhar Alam¹, Abdus S Wahed²

¹Institute of Statistical Research and Training, University of Dhaka, Bangladesh

²Department of Biostatistics and Computational Biology, University of Rochester, New York

Background Randomized controlled trials (RCTs) are the gold standard for estimating causal effects. However, their validity is compromised by noncompliance, particularly when unmeasured confounders influence both compliance and outcomes. In cardiovascular trials, prioritized composite outcomes are common, with the win ratio being a popular method for estimating treatment effects in such cases. The win ratio compares treated and control pairs to determine the winner based on the higher-priority event. If a winner cannot be determined, the secondary event is used for the comparison. Recently developed win ratio methods incorporating propensity scores cannot account for unmeasured confounders, making

them unsuitable in the presence of noncompliance. Common approaches like intention-to-treat (ITT), as-treated (AT), and per-protocol (PP) analyses also fail to provide valid causal estimates when noncompliance occurs.

Methods This study proposes a win ratio estimator that incorporates instrumental variable (IV) to address noncompliance. The idea is to fit a regression model of the treatment received on treatment assigned and extract the residuals. These residuals are then used as a proxy of the unmeasured confounders and then consequently adjusted for to calculate the win ratio. A simple form of the IV win ratio estimator is derived, which does not require any information on the unmeasured confounder.

Results The proposed IV win ratio is compared with simpler alternatives like ITT, AT, and PP using simulations, focusing on bias, standard error, and coverage. Under no treatment effect, AT and PP show high bias and poor coverage even with 5% noncompliance, while ITT and IV remain robust. For strong treatment effects, all estimators perform well under full compliance, but ITT, AT, and PP degrade as compliance decreases. At 85% compliance, their coverage drops to 67%, 23%, and 60%, while IV maintains near-nominal coverage of about 95%. Applying the proposed methods to the JOBS II randomized field experiment data reveals a significant effect of the job training on reemployment and depression among unemployed subjects. Specifically, individuals are 21% more likely to achieve a more favorable outcome (either reemployment or reduced depression) if they received the job-skills training compared to if they just received a booklet.

Conclusion Noncompliance in trials with prioritized composite outcomes is often overlooked. This study introduces an IV win ratio estimator, highlighting its superiority over ITT, AT, and PP win ratios. It provides practical guidance for choosing the most suitable methods for analyzing prioritized composite outcomes in the presence of noncompliance.

Machine Learning, Deep Learning, and AI

Tuesday, 2025-08-26 11:30 - 13:00, Biozentrum U1.101

Chair: Mark van de Wiel

1: Performance Evaluation of Dimensionality Reduction Techniques on High-Dimensional DNA Methylation Data

Kuldeep Kumar Sharma¹, Binukumar B¹, Binu V. S¹, Thirumoorthy Chinnasamy¹, Gokulakrishnan K¹, Saravanan P², Mohan V³

¹National institute of mental health and neurosciences, India

²Warwick Medical School, University of Warwick, UK

³Madras Diabetes Research Foundation (MDRF), Chennai, India

Introduction Most biomedical researchers in the recent past have started recording thousands to millions of features simultaneously on each object or individual, and such data are said to be high-dimensional data. One such field of high-dimensional datasets is epigenetics. Epigenetics is a subbranch of genetics that focuses on inheritable changes in gene activity or function that occur without alterations to the DNA sequence. Datasets obtained from epigenetics are known as DNA methylation (DNAm) data. The methylation datasets include beta values for each cytosine phosphate guanine (CpG), indicating the degree of methylation. A statistical framework for handling such high-dimensional data involves the use of various dimension reduction (DR) techniques. Knowing how each technique performs on DNAm datasets can be helpful in deciding which DR to use.

Objective This communication aims to reduce the dimensionality of DNAm data via various DR techniques, such as principal component analysis (PCA), partial least squares discriminant analysis (PLS-DA), multidimensional scaling (MDS), isometric mapping (ISOMAP) and uniform manifold approximation and projection (UMAP). In addition, there is a comparison of techniques in terms of the retained amount of information, local neighborhood preservation criteria, and global structure-holding approaches.

Methodology: Data for the current study were obtained from the ongoing study MET-BIOWIN. There were 8,62,927 CpG sites, and the dataset consisted of grouping variables for gestational diabetes status (yes/no). Dimension reduction was performed via PCA, PLS-DA, MDS, ISOMAP, and UMAP. Furthermore, the retained amount of information, local neighborhood preservation criteria, and global structure-holding approaches were assessed

via measures such as Shannon's entropy, Spearman's rho, Konig's measure, trustworthiness & continuity, Kruskal's stress score, Sammon's score, and residual variance.

Result PCA performed well in terms of information retention, followed by PLSDA and MDS, while UMAP consistently showed the weakest performance. For local structure preservation, MDS and PCA excelled with high König's measure and trustworthiness, whereas UMAP underperformed significantly. Regarding global structure preservation, MDS and UMAP showed the lowest Kruskal stress scores, indicating a strong fit, while ISOMAP performed the worst. Overall, MDS, and PCA were the most effective, while UMAP lagged across all criteria.

Conclusion Overall, MDS, PCA, and PLSDA emerged as the most robust techniques across multiple metrics, whereas UMAP was consistently less effective.

Keywords DNA methylation data, PCA, PLS-DA, MDS, ISOMAP, UMAP.

2: Deep Generalised Mixed Effects Models: a Novel General Neural Network Structure for Analysing Hierarchical Data

Nina van Gerwen^{1,2}, Dimitris Rizopoulos^{1,2}, Manon Hillegers³, Loes Keijsers⁴, Sten Willemsen^{1,2}

¹Department of Biostatistics, Erasmus University Medical Center, the Netherlands

²Department of Epidemiology, Erasmus University Medical Center, the Netherlands

³Department of Child Psychiatry, Erasmus University Medical Center, the Netherlands

⁴Department of Psychology, Education and Child Studies, Erasmus University Rotterdam, the Netherlands

Background The Experience Sampling Method (ESM) is an intensive longitudinal research design where participants report their thoughts, emotional states and behaviours multiple times a day. ESMs have become increasingly popular to investigate individuals' daily experiences. Our work is motivated by ESM data collected by the Growlt! app. During the COVID-19 pandemic, the app was released to investigate daily mood changes among young adults, give users insight into their emotions and enhance users' resilience. Current procedures to analyse ESM data face various challenges. In particular, ESM data are high-dimensional and exhibit complex correlation structures, which standard statistical techniques (e.g., marginal and mixed effect models) cannot adequately capture. Alternatively, machine learning procedures, such as recurrent neural networks, can be used to model ESM data and accommodate these correlations. However, these procedures face a problem with missing

data. In our motivating dataset, adolescents often stopped using the app due to previous strong feelings of negative emotions. Hence, the implied missing data are of the missing-at-random type that standard machine learning procedures cannot accommodate.

Methods We develop a novel neural network (NN) architecture that generalises mixed effects models to deep learning to overcome these challenges. Our Deep Generalised Mixed Model (DGMM) allows semiparametric and highly flexible modelling of the data's mean and correlation structure with NNs. Classical estimation of mixed models requires integration over the random effects distribution, which is intractable when we estimate the random effects with a NN. Therefore, we use an adaptation of variational autoencoders to estimate the DGMM. By specifying a tractable variational distribution to sample from, we approximate the marginal log-likelihood as an expectation with respect to the variational distribution and the Kullback-Leibler divergence between the variational distribution and the marginal distribution of the random effects, together known as the Evidence Lower Bound. The variational distribution can also be seen as a nonlinear function of the data, which we estimate with another NN. Through this approach, the DGMM is able to accommodate longitudinal outcomes following any generic distribution, scale well to high-dimensional settings and provide valid inference when data is missing-at-random.

Results In the Growlt! app data, the DGMM showed good predictive performance for the multivariate analysis of longitudinal outcomes. A simulation study of the DGMM also showed good performance in various settings.

Conclusion We have implemented the DGMM in Python using Keras, and are developing a wrapper function for R users.

3: Adapting Transformer Neural Networks for Longitudinal Data with few Time Points

Kiana Farhadyar, Harald Binder

Institute of Medical Biometry and Statistics, Germany

When simultaneously assessing several characteristics of individuals over time, there might be a complex pattern of relations between these. While autoregressive models can be useful in such a setting when there is a smooth continuous temporal pattern, they rely on fixed lag structures and struggle with a small number of time points. The attention-based weighting scheme in the transformer neural network architectures might be better suited for discontinuous patterns, as it can assign weights to different time points, but so far, it has been limited

to rather large datasets due to a large number of parameters. Therefore, we created a considerably simplified architecture that still maintains the key characteristics of transformers. This also allowed us to design a statistical testing approach for identifying context characteristics that steer the effect of other characteristics. This is complemented by a visualization approach for illustrating the pairwise relevance of characteristics. We illustrate our proposed technique using both simulated data and real-world data comprising self-reported stressors from a longitudinal resilience assessment study. There, prediction performance is seen to improve over classical regression approaches. In addition, the statistical testing approach uncovers the underlying patterns in the simulation study and highlights significant features in the real data that align with the mental health dynamics.

4: Distinguishing Subgroup and Site-Specific Heterogeneity in Multi-Site Prognostic Models using a Neural Network Representation

Max Behrens^{1,2}, Daiana Stoltz³, Eleni Papakonstantinou³, Janis M. Nolde⁴, Gabriele Bellerino⁵, Moritz Hess^{1,2}, Harald Binder^{1,2}

¹Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Stefan-Meier-Straße 26, 79104 Freiburg, Germany

²Freiburg Center for Data Analysis and Modeling, University of Freiburg, Ernst-Zermelo-Straße 1, 79104 Freiburg, Germany

³Clinic of Pneumology, Medical Center – University of Freiburg, Faculty of Medicine, University of Freiburg, Killianstrasse 5, 79106 Freiburg, Germany

⁴Department of Nephrology, Faculty of Medicine and Medical Center, University of Freiburg, Hugstetter Strasse 55, 79106 Freiburg, Germany

⁵University of Freiburg, Department of Mathematical Stochastics, Ernst-Zermelo-Straße 1, 79104 Freiburg, Germany

Introduction Multi-site clinical studies often struggle with heterogeneity in the effects of patient characteristics on outcomes. This variability may arise not only from inherent differences between (unknown) patient subgroups but also from systematic variations across sites, including differing subgroup compositions or clinical practices. Often, we assume homogeneity or rely on site-level adjustments through interaction terms or random effects models, not covering the full spectrum of heterogeneity. We propose a more flexible approach leveraging a low-dimensional representation, obtained via neural networks, to quantify different sources of heterogeneity. Specifically, the aim is to distinguish between patient subgroup differences that are invariant across sites and site-specific differences.

Methods Our approach employs an autoencoder to learn a low-dimensional representation of

the high-dimensional patient characteristic space, preserving essential patterns while reducing noise and redundancy. We integrate prognostic modelling directly within this learned latent representation. For each patient, we fit a localized regression model—weighting nearby patients based on their proximity in the latent space—to obtain patient-specific coefficient estimates that capture local prognostic variations. By examining the distribution of these patient-specific coefficients, we quantify overall heterogeneity for each latent dimension. Further, we can disentangle site effects by analysing patterns in these coefficients across sites. The entire method, from autoencoder training to localized regression, is trained end-to-end. This joint optimization ensures the latent representation not only preserves patterns in patient characteristics but also local variations in prognostic effects, making heterogeneity patterns more discernible.

Results We illustrate our method with a multi-site dataset of patients with chronic obstructive pulmonary disease, investigating the heterogeneity in prognostic factors for disease progression. Within individual sites, distinct subgroups were found that differed in their coefficient estimates. Across sites, we identified both site-invariant subgroups and site-specific subgroups, likely reflecting differences in patient demographics or clinical practices. To aid interpretations, we provide a visualization approach for translating these patterns back to the level of patient characteristics.

Conclusion Our autoencoder-based approach provides a tool for quantifying and decomposing heterogeneity in multi-site clinical data. By learning a latent representation optimized for both data reconstruction and preservation of local prognostic effects, we can effectively identify and differentiate site-specific and site-invariant sources of heterogeneity. This method allows for a more nuanced understanding of prognostic factors, moving beyond global estimates and facilitating the development of more robust and personalized models.

5: Unraveling Breast Cancer Genetic Risk in Chinese Women: Integrating GWAS, Fine-Mapping, and Machine Learning in the China Kadoorie Biobank

Shizhe Xu¹, Christiana Kartsonaki¹, Kuang Lin¹, Kyriaki Michailidou²

¹University of Oxford, United Kingdom

²The Cyprus Institute of Neurology and Genetics, Cyprus

Background / Introduction Genome-wide association studies (GWAS) have identified approximately 200 genomic regions containing common genetic variants associated with breast cancer risk. However, their target genes remain uncertain mainly due to linkage disequilibrium (LD) and the prevalence of variants in non-coding regions. To address this, fine-mapping

methods have been introduced to pinpoint the most likely causal variants from a set of credible candidate variants and identify target genes. Most previous GWAS and fine-mapping studies have primarily focused on European-ancestry individuals. Given differences in genetic architecture and environmental exposures between Asian and European populations, our study aims to conduct GWAS on Chinese women and perform fine-mapping with summary statistics to uncover additional association signals and candidate susceptibility genes for breast cancer. Furthermore, we integrate machine learning models into the fine-mapping process.

Methods First, we performed a GWAS on 57,660 Chinese women from the China Kadoorie Biobank using two software packages, SAIGE and REGENIE. Second, to understand how these packages handle complex living regions in China, we analysed specific loci and explicitly compared their approaches to computing relatedness and performing association testing by a Firth logistic regression model or a linear mixed model. Third, to distinguish true causal variants from significant signals, we applied fine-mapping methods such as SuSiE-RSS and PolyFun to our generated summary statistics. Fourth, to investigate computational trade-offs, we conducted fine-mapping on individual-level data and summary statistics to evaluate loss in accuracy. Finally, we applied a sequence-based deep learning model to assign functional annotations to variants in non-coding regions and incorporated a supervised learning approach, such as random forest.

Results Summary statistics, Manhattan plots, QQ plots and LD score regression files were generated. Among Chinese women, several genetic loci associated with breast cancer were identified. The signals detected by SAIGE were slightly more significant than those identified by REGENIE due to differences in their underlying algorithms and correction thresholds. A comparison was conducted between fine-mapping results derived from individual-level data and summary statistics. A systematic evaluation was conducted to assess whether functional annotations and supervised learning enhance fine-mapping accuracy.

Conclusion This study provides new insights into breast cancer genetics based on data from Chinese women. The comparison between REGENIE and SAIGE enhances our understanding of their strengths and limitations. This study also visualises the differences between fine-mapping using individual-level data and summary statistics. Finally, a machine learning-based framework in fine-mapping paves the way for more explicit analysis.

Meta-Analysis 2

Tuesday, 2025-08-26 11:30 - 13:00, ETH E27

Chair: Keith R Abrams

1: How to Quantify Between-Study Heterogeneity in Single-Arm Evidence Synthesis

Ulrike Held¹, Lea Bührer^{1,2}, Beatrix Latal³, Stefania Iaquinto¹

¹Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Switzerland

²Centre for Computational Health, Institute of Computational Life Sciences, Zurich University of Applied Sciences (ZHAW), Wädenswil, Switzerland

³Child Development Center, University Children's Hospital, University of Zurich, Switzerland

Background Random-effects meta-analysis models account for between-study heterogeneity by estimating and incorporating the heterogeneity variance parameter τ^2 . Different estimators for τ^2 have been proposed, but no widely accepted guidance exists on when to best use which meta-analysis variance estimator. In the context of meta-analysis of single-arm observational studies, studies with unique challenges, such as large variability in outcomes, sparse data, and high methodological heterogeneity, systematic evaluations and comparisons of the different heterogeneity variance estimators are lacking. A neutral comparison simulation study was conducted to represent typical meta-analysis scenarios for continuous and binary outcomes in a single-arm meta-analysis setting. Furthermore, a non-systematic literature review was conducted, and the methods were applied to a case study involving infants with congenital heart disease (1).

Methods Seven different estimators of τ^2 were preselected based on their use in clinical research and their availability in the R programming environment. Their performance was assessed in terms of mean bias, mean squared error, and the proportion of estimates equal to zero. Additionally, coverage and bias-eliminated coverage were evaluated using Wald and Hartung-Knapp confidence intervals. Prediction intervals were additionally calculated. In a non-systematic literature review, we assessed which meta-analysis methods are currently used in high-ranked medical journals.

Results Our neutral comparison simulation study showed imprecision across all heterogeneity variance estimators, particularly in meta-analyses with a small number of studies or when analysing binary outcomes with rare events. Many heterogeneity variance estimators fre-

quently produced zero heterogeneity estimates, even in the presence of heterogeneity. Notably, while the estimated overall effects remained relatively robust, prediction intervals varied substantially across methods. Additionally, our literature review indicated a low level of statistical literacy regarding heterogeneity variance estimators in single-arm meta-analyses, with over half of the reviewed studies failing to report the estimator used. A preprint of our study is available (2).

Conclusion We conclude that relying on a single heterogeneity variance estimator is not appropriate for single-arm meta-analysis of observational studies. Instead, we recommend using multiple estimators in a sensitivity analysis, especially when evaluating prediction intervals.

References (1) Feldmann, M., Bataillard, C., Ehrler, M., Ullrich, C., Knirsch, W., Gosteli-Peter, M. A., Held, U., Latal, B. (2021). Cognitive and Executive Function in Congenital Heart Disease: A Meta-analysis. *Pediatrics*, 148(4), e2021050875. <https://doi.org/10.1542/peds.2021-050875>

(2) Iaquinto, S., Feldmann, M., Latal B. et al. How to quantify between-study heterogeneity in single-group evidence synthesis? - It depends!, PREPRINT (Version 1) available at Research Square [<https://doi.org/10.21203/rs.3.rs-2450618/v1>]

2: Bayesian Random-Effects Meta-Analysis with Empirical Heterogeneity Priors for HTA Applications in the Situation of Very few Studies

Ralf Bender¹, Jona Lilienthal¹, Sibylle Sturtz¹, Christian Röver², Tim Friede²

¹Department of Medical Biometry, Institute for Quality and Efficiency in Health Care (IQWiG), Cologne, Germany

²Department of Medical Statistics, University Medical Center Göttingen, Göttingen, Germany

Background In Bayesian random-effects meta-analysis, the use of (weakly) informative heterogeneity priors is of particular benefit in the case of very few studies, a situation often encountered in HTA applications [1]. Empirical heterogeneity priors derived from Cochrane reviews are available but it is unclear whether these are adequate for HTA applications [5]. Different heterogeneity priors have been proposed in the literature as well as methods to derive prior distributions from empirical meta-analyses [3-5].

Methods We collected all relevant meta-analyses from IQWiG reports for the period 2005 to 2021. We considered the effect measures SMD for continuous data, HR for time-to-

event data, and OR and RR for binary data. The heterogeneity parameter was re-estimated by applying random-effects meta-analyses with the Knapp-Hartung and the Paule-Mandel method. The hierarchical Bayesian model proposed by Röver et al. [4] was applied to derive empirical heterogeneity priors for the different effect measures. We compared these with previous proposals for heterogeneity priors [3,5] and compared the meta-analytic results of the Bayesian approach with those from the former IQWiG approach for evidence synthesis in the case of very few studies.

Results Empirical heterogeneity priors based on the half-normal distribution are derived, which have more distributional weight on smaller heterogeneity values than previous suggestions. Evidence synthesis based on the new heterogeneity priors more frequently allows for a quantification of the treatment effect than the former IQWiG approach [2].

Conclusions The new heterogeneity priors are suitable for the application of Bayesian random-effects meta-analyses with very few studies in the HTA framework.

References [1] Bender, R. et al. (2018): Methods for evidence synthesis in the case of very few studies. *Res. Syn. Methods* **9**, 382–392.

[2] Lilienthal, J. et al. (2024): Bayesian random-effects meta-analysis with empirical heterogeneity priors for application in health technology assessment with very few studies. *Res. Syn. Methods* **15**, 275–287.

[3] Röver, C. et al. (2021): On weakly informative prior distributions for the heterogeneity parameter in Bayesian random-effects meta-analysis. *Res. Syn. Methods* **12**, 448–474.

[4] Röver, C. et al. (2023): Summarizing empirical information on between-study heterogeneity for Bayesian random-effects meta-analysis. *Stat. Med.* **42**, 2439–2454.

[5] Turner, R.M. et al. (2015): Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis. *Stat. Med.* **34**, 984–998.

3: Assessing Inconsistency in Flexible Meta-Regression using Network Meta Analysis

Marc Angelo Parsons, Andrea Benedetti, Russell Steele

McGill University, Canada

Background Medical research is often interested in changes in health outcomes over time. These are called trajectories. While flexible regression methods such as spline or fractional polynomial models are well developed for single studies, the case is less clear in the case of trajectory meta-analysis (MA) of multiple primary studies in the presence of differences in outcome assessment patterns between studies.

Network meta-analysis (NMA) simultaneously estimates multiple pairwise effect differences between a batch of treatments. This method has well-defined assumptions and an accepted framework for evaluating results. The fundamental assumption of consistency between direct and indirect comparisons can be quantified in several ways. By measuring consistency in a longitudinal context, we could assess the impact of heterogeneous outcome assessment patterns.

Methods We present a novel application of NMA to the estimation of longitudinal trajectories: Discrete Time NMA (DTNMA). Briefly, DTNMA considers pairwise comparisons between discrete timepoints rather than treatments, leveraging both direct and indirect evidence to plot out an estimated trajectory. This paper outlines the underlying methodology, including assumptions, of DTNMA. Two case studies are presented: one using the motivating dataset presented above and another employing a systematic review of depression scores measured during the COVID-19 pandemic.

Results In the first case study, we found moderate evidence for a decline in depression scores over the perinatal period. In the second case study, we found no evidence for change in depression scores over the two years beyond the start of the COVID-19 pandemic. However, for both case studies, we found evidence for high levels of heterogeneity and inconsistency between studies due to differing outcome assessment patterns using the DTNMA models.

Conclusion The novel method presented in this paper (DTNMA) provides researchers with a method to assess inconsistency in MA of trajectories due to irregular outcome assessment patterns between included studies. In standard flexible meta-regression, it is not clear how to measure the effect of this issue on model results. We have shown that using an NMA approach can provide a possible pathway to resolve this problem. In the case studies presented, high levels of heterogeneity and inconsistency in the patterns of outcome assessment may limit the final conclusions of model results. We have demonstrated how inconsistency in an NMA can provide researchers with insight into the previously unconsidered issue of differing outcome assessment patterns between included studies.

4: Illuminating the Assumptions of Meta-Regression in Treatment Networks

Nana-adjoa Kwarteng¹, Theodoros Evrenoglou¹, Adriani Nikolakopoulou^{1,2}

¹Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Centre, University of Freiburg, Freiburg im Breisgau, Germany

²Department of Hygiene, Social-Preventive Medicine and Medical Statistics, School of Medicine, Aristotle University of Thessaloniki, Thessaloniki, Greece

Background / Introduction Network meta-analysis (NMA) is a common statistical method used to synthesize evidence from multiple studies and enable the simultaneous comparison of multiple competing treatments for a given condition. When meta-analysts wish to explore remaining heterogeneity, Network meta-regression (NMR) provides a valuable extension to NMA by adjusting treatment effect estimates based on study-level characteristics in an attempt to further explain heterogeneity across studies. Despite its usefulness, NMR adoption as a technique has been limited due to challenges regarding conceptual issues, implementation accessibility, and limitations in sparse data settings. NMR model coefficients follow independent, exchangeable, or common assumptions, each with or without consistency in treatment comparisons. However, choosing between these models is complex, as different modeling assumptions can lead to varying results and interpretations depending on factors like model fit and data availability. Additionally before fitting NMR models, meta-analysts must examine the relevance of data directionality, where certain study characteristics (e.g., sponsorship) systematically influence treatment effect estimates, introducing potential bias.

Methods To address these challenges, we introduce frequentist tools and a graphical toolkit for NMR models, extending the landscape of implementation tools available to analysts. Additionally, we investigate different modeling assumptions and explore their properties under various scenarios. Based on the properties of the available model assumptions, we provide recommendations for implementation of NMR when considering data availability and the research question of interest. We also provide guidance for the interpretation of treatment-by-covariate interactions in relation to their underlying NMR assumptions.

Results We illustrate our methods by comparing different NMR modeling assumptions in a network of 10 diabetes treatments, and simulations with various data generation scenarios. Detailed examination of the output highlights the importance of directionality in models that do not impose consistency in the treatment-by-covariate interactions. Furthermore, results emphasize complexities in scenarios with small subgroups with a lack of sufficient variation in the observed covariates.

Conclusion This work elucidates some NMR assumptions, the importance of the network structure, and the importance of small data for NMR. The introduced tools can streamline the implementation of NMR, facilitating the exploration of sources of heterogeneity and inconsistency in NMA and expanding tools available to meta-analysts of evidence synthesis.

5: Transitivity in Network Meta-Analysis: A Formal Framework and Practical Implications

Noosheen Rajabzadeh Tahmasebi¹, Ian White², Georgia Salanti³, Efthimiou Orestis^{3,4}, Adriani Nikolakopoulou^{1,5}

¹Institut für Medizinische Biometrie und Statistik (IMBI), Universitätsklinikum freiburg (Germany)

²MRC Clinical Trials Unit, University College London (UK)

³Institute of Social and Preventive Medicine (ISPM), University of Bern (Switzerland)

⁴Berner Institut für Hausarztmedizin (BIHAM), University of Bern (Switzerland)

⁵Laboratory of Hygiene, Social and Preventive Medicine and Medical Statistics, Aristotle University of Thessaloniki (Greece)

Network meta-analysis (NMA) synthesizes evidence for multiple treatments, relying on assumptions of transitivity and consistency. Transitivity has been widely discussed, and many interpretations and practical considerations have been suggested. However, the formal definition of transitivity and its relation with consistency remain ambiguous. We propose a clear definition of transitivity, viewing it as a property of counterfactual treatment effects derived from joint randomizability. In particular, we define transitivity as the equality of true effects between treatments X and Y across studies, even when a study does not include X or Y. Subsequently, we show how consistency, defined as the agreement between different sources of evidence, can be derived as a consequence of this definition. We show how common interpretations of transitivity relate to its formal definition. We then link the transitivity assumption with assumptions about missing data mechanisms. Specifically, we demonstrate that transitivity is equivalent to the assumption of missing completely at random and that a weaker assumption, missing at random, suffices for valid NMAs when paired with likelihood-based analyses. Our findings highlight key properties of the assumptions of NMA and call for careful examination of these assumptions (e.g., through examining the distributions of effect modifiers) to enhance the robustness of evidence synthesis.

Meta Science 2

Tuesday, 2025-08-26 11:30 - 13:00, ETH E23

Chair: Carsten Oliver Schmidt

1: Developing Communication Skills in Biostatistical Consulting and Collaboration

Karen Lamb, Sabine Braat, Julie Simpson

University of Melbourne, Australia

Communication is a key skill in biostatistical consultancy and yet it is something that is not emphasised in our training. One of the most challenging things about embarking on a career in statistical consulting is learning how to say “no” or, more often, “not like that” or “not right now”. Although it is important to say “yes” to many projects for strategic or financial purposes, to develop new expertise, or due to interest in the topic, there are also many reasons and ways to say no. Of particular concern is feeling you have to say “yes” to something that is clearly fraudulent or unethical. In a 2018 US study of 390 consulting biostatisticians, findings showed that researchers often make “inappropriate requests” of statisticians, ranging from being asked to fake statistical significance, to changing or removing data¹. Clearly, it is important that biostatisticians feel able to say no, but this can be difficult, particularly for early career biostatisticians who feel under pressure to acquiesce to the requests of their seniors.

The Methods and Implementation Support for Clinical and Health research (MISCH) Hub provides support in key aspects of clinical and health research to researchers at the University of Melbourne and affiliated hospital partners, including biostatistics, health economics and co-design. As co-Head of biostatistics for MISCH, one aspect of my role is to assist biostatisticians to develop confidence in how to effectively communicate with researchers. This includes helping develop the ability to negotiate with collaborators. In this presentation, I will describe situations in which a “no”, “not like that” or “not right now” has been the approach I have opted for and how I negotiated alternative solutions for the collaborator. The case study examples range from grant applications with inappropriate study designs, convincing researchers to move on from ANOVA, incorporating the estimand framework in trials, and negotiating achievable deadlines with collaborators. In addition, I will describe some approaches we use within MISCH to support the development of communication skills in our team to help them in their roles as biostatistical consultants within academia.

REFERENCES

¹Wang et al. Researcher Requests for Inappropriate Analysis and Reporting: A U.S. Survey of Consulting Biostatisticians. *Ann Intern Med* 2018;169(8):554-558.

2: Are We “essential” or “not Needed”? Varying Perceptions of Statisticians’ Value, a National Study of Human Research Ethics Committees

Adrian Barnett¹, Nicole White¹, Taya A Collyer²

¹Australian Centre for Health Services Innovation and Centre for Healthcare Transformation, Queensland University of Technology

²National Centre for Healthy Ageing, Monash University, Australia

Background Inappropriate study design and statistical analysis leads to research waste, and wasted participant effort. Proposed clinical research is generally reviewed by an ethics committee, and ethical review represents a key opportunity for inappropriate designs to be identified and remedied. However, the availability and quality of statistical advice for ethics committees is inconsistent, due to a lack of standardisation. In Australia, the proportion of committees with access to a formally-qualified statistician is unknown, and the beliefs and attitudes regarding the role of statisticians in ethical review are unclear.

Methods To explore these issues, we approached all human research ethics committees in Australia, to complete an online survey with open and closed questions about the role of statistical advice in the committee’s work. Structured questions were analysed descriptively, and text responses to open questions were analysed qualitatively, via thematic analysis with both inductive and deductive code-sets.

Results Sixty percent of committees reported access to a statistician, either as a full committee member or as a non-member who could be consulted, but the reduced to 35% when accounting for formal statistical qualifications. Many committees reported relying on “experienced” or “highly numerate” researchers in place of qualified statisticians, as they view general research experience and advanced statistical training as equivalent.

Committees without access to statisticians tended to locate responsibility for study design with other parties, including researchers, trial sponsors, and institutions. Some committee chairs viewed formal statistical input as essential to the work of their committee; however,

amongst those who viewed statistical advice as unimportant or unnecessary, there was a widespread belief that statistical review is only applicable to particular kinds of studies, and that “simple”, observational or “small” studies do not merit statistical review.

Conclusion We encountered dramatic and surprising variance in practice and attitudes towards the role of statisticians on human research ethics committees. Concerningly, qualitative analysis revealed that some practices and attitudes are underpinned by beliefs about statistics and statisticians which are demonstrably incorrect. The number of research studies receiving approval without statistical review in Australia is concerning, risking studies that in the best-case waste resources, and in the worst-case cause harm due to flawed evidence.

3: An Empirical Assessment of the Cost of Dichotomization

Erik van Zwet¹, Frank Harrell², Stephen Senn³

¹Leiden University Medical Center, The Netherlands

²Vanderbilt University Medical Center, Tennessee, US

³Edinburgh, UK

Background We consider two-arm parallel clinical trials. It is well known that binary outcomes are less informative than continuous (numerical) ones. This must be compensated by larger sample sizes to maintain sufficient power.

Methods If the continuous outcome has the normal distribution, then the standardized mean difference (SMD) and the probit transformation of the dichotomized outcome are both estimates of Cohen's d. We use this equivalence to study the loss of information due to dichotomization. We have used 21,435 unique randomized controlled trials (RCTs) from the Cochrane Database of Systematic Reviews (CDSR). Of these trials, 7,224 (34%) have a continuous (numerical) outcome and 14,211 (66%) have a binary outcome. We find that trials with a binary outcome have larger sample sizes on average, but also larger standard errors and fewer statistically significant results.

Results We conclude that researchers do tend to increase the sample size to compensate for the low information content of binary outcomes, but not nearly sufficiently.

Conclusion In many cases, the binary outcome is the result of dichotomization of a continuous outcome which is sometimes referred to as “responder analysis”. In those cases, the loss of information is avoidable. Burdening more subjects than necessary is wasteful, costly and

unethical. We provide a method to calculate by how much the sample size may be reduced if the outcome would not be dichotomized. We hope that this will guide researchers during the planning phase. We also provide a method to calculate the loss of information after a responder analysis has been done. We hope that this will motivate researchers to abandon dichotomization in future trials.

4: Reporting Completeness in Conventional and Machine Learning-Based COVID-19 Prognostic Models: A Meta-Epidemiological Study

Ioannis Partheniadis, Persefoni Talimtzi, Adriani Nikolakopoulou, Anna Haidich

Aristotle University of Thessaloniki, Greece

Background The rapid publication of prognostic prediction models for COVID-19 presents an opportunity to study the evaluation of the reporting completeness, to elucidate the potential of their clinical applicability. This study assesses and compares the reporting completeness of conventional and machine learning-based models using the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) [1] and its AI extension (TRIPOD+AI) [2].

Methods We included studies reporting the development, and internal and external validation of prognostic prediction models for COVID-19 using either conventional or machine learning-based algorithms. Literature searches were conducted in MEDLINE, Epistemonikos.org, and Scopus (up to July 31, 2024). Studies using conventional statistical methods were evaluated under TRIPOD, while machine learning-based studies were assessed using TRIPOD+AI. Data extraction followed TRIPOD and TRIPOD+AI checklists, measuring adherence per article and per checklist item. The protocol was registered on the Open Science Framework (<https://osf.io/kg9yw>).

Results We identified 53 studies reporting 71 prognostic models. On average, studies using conventional models adhered to 38.1% (SD: ± 10.4) of the TRIPOD checklist, while machine learning-based studies adhered to 28.37% (SD: ± 8.91) of TRIPOD+AI. No study fully adhered to abstract reporting requirements, and few included an appropriate title (29.0%, 95% CI: 16.1–46.6 for TRIPOD; 13.6%, 95% CI: 4.8–33.3 for TRIPOD+AI). Notably, no study fully reported a sample size assessment. Reporting of methods and results sections was poor across both frameworks. Overall, adherence to TRIPOD and TRIPOD+AI guidelines was generally low, with machine learning-based models showing significantly lower overall adherence (28.4% vs. 38.1%) ($p < 0.001$). The lower adherence to the TRIPOD+AI statement was somewhat expected, as these guidelines were published in April 2024 [2], two years after

the most recent study included in this analysis.

Conclusion Reporting completeness was inadequate for both conventional and machine learning-based models, with critical omissions in model specifications and performance metrics. Strengthening adherence to reporting guidelines is essential to enhance research transparency, prevent research waste, and improve clinical utility.

[1] Collins GS, Reitsma JB, Altman DG *et al.* Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Med* **13**, 1 (2015). doi.org/10.1186/s12916-014-0241-z.

[2] Collins GS, Moons KGM, Dhiman P, *et al.* TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **385**, e078378 (2024). doi:10.1136/bmj-2023-078378.

5: How to Improve Data Quality Reporting: a Real-World Data Example using Registry Data

Elena Salogni¹, Thomas J. Musholt², Elisa Kasbohm¹, Stephan Struckmann¹, Carsten Oliver Schmidt¹

¹Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany

²Section of Endocrine Surgery, Department of General, Visceral and Transplantation Surgery, University Medical Centre, Johannes Gutenberg-University Mainz, Mainz, Germany

Background / Introduction Knowing the properties and quality of research data is essential for credible research findings. However, achieving a comprehensive overview can be difficult if the data have been obtained outside a highly controlled environment. This is often the case with real-world data collections, such as clinical registries. Ideally, registry administrators would provide extensive data overviews in a standardized manner. However, this is not routinely the case. This talk presents an information workflow to transparently check data with the R package dataquieR [1]. We illustrate how this approach provides clear and actionable insights on data quality issues using the Eurocrine registry, a European registry that contains data on the diagnosis and surgical treatment of endocrine tumors and diseases.

Methods The Eurocrine registry contains data from more than 170.000 patients and almost 150 clinics across 17 European countries. Assessed data cover preoperative, operative, post-operative, and follow-up data. Metadata defining expectations for the data were assembled in an Excel worksheet. With a single command call, dataquieR produces data quality reports

comprising descriptive statistics and up to 24 data quality indicators related to data formatting, missing data, range violations, outliers, contradictions, cluster effects, and time trends. All findings are assembled in an extensive html report and can be used for subsequent data corrections.

Results Due to the numerous calculations, the computational time (Windows 10, 128GB RAM, Core i7-12 cores) was approximately 10-12 hours. The report reveals issues related to data formatting errors, completeness, and data correctness. All of the mandatory variables are almost 100% complete. A few inadmissible values (<2%) and implausible data (e.g., postoperative serum calcium level) can be observed. Contradictions checks have a percentage of violation below 2%.

Conclusion The presented approach is generic and can be applied similarly to other data sources. In the Eurocrine application example, previously unknown issues were discovered despite existing measures to secure a high data quality. Findings can be of relevance for decisions on statistical analyses but also as part of a data monitoring to improve, where possible, the quality of the data base.

- [1] Struckmann S., et al. (2024). dataquieR 2: An updated R package for FAIR data quality assessments in observational studies and electronic health record data. JOSS; 9(98):6581. 10.21105/joss.06581.

Statistical Methods in Epidemiology

Tuesday, 2025-08-26 11:30 - 13:00, ETH E21

Chair: Stefania Galimberti

1: Federated Inference Methods for Estimation and Comparison of Standardized Mortality Ratios

Zoë D. van den Heuvel, Bas de Groot, Marianne A. Jonker

RadboudUMC, Netherlands, The

One way to benchmark quality of care in emergency departments of medical centers is by comparison of quality indicators like the Standardized Mortality Ratio (SMR), in order to improve patient outcomes where possible. The SMR is defined as the ratio of the observed and the expected mortality rate. To be able to estimate the SMRs that account for the different patient populations in the medical centers, it is necessary to combine the data from different medical centers into a single database. However, sharing data across medical centers is in practice challenging due to regulatory and privacy problems.

Recently, Bayesian Federated Inference (BFI) was introduced to construct from local inferences in separate medical centers what would have been inferred had the data sets been merged [1]. In this methodology, the estimates and statistical power of a combined database can be obtained, without actually constructing this database. The aim of this research is to apply the BFI methodology to real world emergency department data from multiple medical centers, in order to estimate and compare SMRs, adjusted for case-mix. In the presentation we explain how the BFI methodology can be applied to achieve this, without combining data from the different centers.

[1] Jonker, M. A., Pazira, H., & Coolen, A. C. (2024). Bayesian federated inference for estimating statistical models based on non-shared multicenter data sets. *Statistics in Medicine*, 43(12), 2421-2438.

2: A Mixture Model for Subtype Identification: Application to CADASIL

Sofia Kaisaridi¹, Juliette Ortholand¹, Caglayan Tuna², Nicolas Gensollen¹, Sophie Tezenas du Montcel¹

¹ARAMIS, Sorbonne Université, Institut du Cerveau-Paris Brain Institute-ICM, CNRS, Inria, Inserm, AP-HP, Groupe Hospitalier Sorbonne Université, Paris, France

²Inria, Université Paris Cité, Inserm, HeKA, F-75015 Paris, France

Background / 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. However, these techniques often assume that individuals form a homogeneous cluster, thus ignoring possible subgroups within the population. Taking into account different subtypes of progression, while estimating the average course of the disease, is an important task, particularly for diseases with poorly understood underlying mechanisms. Cerebral Autosomal Dominant Arteriopathy with Subcortical Infarcts and Leukoencephalopathy (CADASIL) has been shown to be such an example, with a post-hoc classification study revealing two different subtypes¹ depending on the spatiotemporal variability. The aim of this study is to extend an existing mixed effects model, to identify different clusters of disease progression at the time when the estimation task is performed.

Methods We integrate a probabilistic mixture framework, into an existing non-linear mixed-effect model used for disease course mapping, implemented in the open-source python library Leaspy. In this framework, inter-individual variability is captured through three spatiotemporal parameters: the disease onset, the pace of progression and the ordering of the symptoms. We add a new layer atop the hierarchical structure of the model, assuming that the random effects are coming from a mixture of normal distributions, with their respective probabilities of occurrence. The joint estimation of clusters and model parameters is performed using a mixture Monte-Carlo Markov chain stochastic approximation Expectation Maximisation (M-MCMC SAEM) algorithm.

Results We show that our model successfully recovers the ground truth parameters from synthetic data, with reduced bias comparing to the naïve solution of post-hoc clustering of individual parameters from a one-class model. Our application to CADASIL disease data allows the unsupervised identification of disease subtypes, involving all aspects of the modelled spatiotemporal variability.

Conclusion We proposed a mixture model, as an extension of the disease course mapping model (Leaspy), to properly identify the underlying subgroups of progression based on the individual parameters. This work aims to contribute to the complex challenge of uncovering the heterogeneity arising from different subtypes in chronic diseases.

- 1.Kaisaridi S., Herve D., Jabouley A., Reyes S., Machado C., Guey S., Taleb A., Fernandes F., Chabriat H. et Tezenas du Montcel S. (2025). Determining Clinical Disease Progression in Symptomatic Patients With CADASIL. *Neurology*, 104(1), e210193

3: Performance of a Residual-Based Algorithm Aiming at Identifying Response Shift at the Item Level using Rasch Models: a Simulation Study

Yseulys Dubuy¹, Victor Rechard¹, Véronique Sébille^{1,2}

¹Nantes Université, U1246 SPHERE "methodS in Patient-centered outcomes & HEalth Re-sEarch", France

²Nantes Université, CHU Nantes, Methodology and Biostatistics Unit, Nantes, France

Background One of the main challenges when analyzing longitudinal Patient-Reported Outcomes (PROs) data is that items interpretation in questionnaires designed to measure the PRO of interest (e.g., fatigue, anxiety) can change over time. As a result, the observed change in patient responses reflects both the change in the PRO itself and the changes in the items interpretation, referred to as Response Shift (RS). RS can arise due to the experiencing challenging health events (e.g., salient events or living with a progressive chronic condition). Examining RS at the item level is crucial, as it may provide valuable insights into patients' experiences, notably regarding their potential psychological adjustment to challenging health events. Furthermore, ignoring RS can interfere with the inferences made from PRO data.

Methods Inspired by Andrich & Hagquist's work on differential item functioning (a phenomenon related to RS), we developed an item-level RS detection procedure based on the residuals of a random-effect Partial Credit Model. Specifically, we aimed at determining whether the residuals distribution is associated with time, as an operationalization of RS. The performance of this newly developed procedure was evaluated through a simulation study involving two measurement occasions. Different scenarios were considered, in which varied: the number of items and response categories of the questionnaire, the sample size, the mean change in the PRO levels over time, the presence/absence of RS, the RS effect size, and the proportion of items affected by RS (i.e. proportion of RS items). The performance of the procedure was assessed based on: (1) the rates of false and correct detection of RS (in scenarios without/with simulated RS, respectively), (2) RS recovery (whether the procedure identified the items on which RS was simulated) and (3) the bias when estimating the change in the PRO levels over time, after accounting for RS effects evidenced.

Results The rate of false detection of RS should be lower than 5%, due to correction for multiple testing. The rate of correct detection of RS will likely be influenced by sample size, RS effect size, and proportion of RS items. Specifically, higher sample sizes, RS effect sizes, and proportion of RS items are expected to increase correct RS detection rate.

Conclusions Detecting and accounting for RS is crucial to avoid suboptimal healthcare decision-making. The newly developed procedure might outperform existing item-level RS detection methods. Moreover, this approach can be extended to integrate covariates, helping

to identify RS determinants.

References Andrich&Hagquist, <https://doi.org/10.1186/s12955-017-0755-0>

4: Lifecourse Modelling and Time-Varying Covariates: Empirical Application and Simulation Study for Novel Method

Solomon Beer¹, Sherief Eldeeb², Erin Dunn², Andrew Simpkin¹, Andrew Smith³

¹University of Galway, Ireland

²Purdue University, Indiana

³University of the West of England, United Kingdom

In prospective cohort studies scientists collect many repeated measures, including exposures and outcomes, that are highly correlated over time. Where an exposure is collected repeatedly, interest often lies in determining whether timing has a differential effect on a later outcome. However, few such studies consider the effect of time-varying covariates (TVC) which may impact associations identified.

One approach for such data is the Structured Life Course Modeling Approach (SLCMA), where users select between temporal hypotheses of exposure specified a priori, selecting the hypothesis that explains the most variation in outcome. However, traditionally SLCMA has not accounted for time-varying covariates. We present a modified version of this approach - direct and mediated effects (DME) SLCMA - which corrects for TVC by adjusting on a per hypothesis basis only for covariates measured before the time of each respective temporal hypothesis. A covariate could be a confounder or mediator depending on the temporal hypotheses tested.

In a simulation study, informed by empirical data from the Drakenstein Child Health Study (DCHS), we compared the existing and modified SLCMA and found several scenarios where TVC had a major effect on selecting an incorrect hypothesis. In particular, where the covariates have a greater causal effect on the following exposure than exposures have on the following covariate, and especially when no direct exposure effects are present. Only in scenarios with no indirect effects did SLCMA always select the correct hypothesis, indicating the importance of correcting for relevant TVC when confounding is plausible.

As an application of this method, we use DME SLCMA on DCHS data to investigate the

importance that timing of exposure to maternal psychopathology (repeatedly measured at birth, age 4 and 8) has on childhood depression measured at age 8 whilst correcting for time-varying socioeconomic position (parental assets measured at birth, age 4 and 8). Exposure to childhood adversity is a potent risk factor for later negative mental health outcomes, with growing evidence that the developmental timing of adversity exposure is important. Exposure at age 8 is found to have the strongest effect on childhood depression, even after accounting for time-varying covariates.

To our knowledge, this study is the first to account for time-varying covariates in a lifecourse modelling approach. Our results are useful for future studies where it is essential to adjust for time-varying covariates to prevent incorrect and biased estimates, such as when the exposure of interest acts only indirectly on the outcome.

5: Emulating Hypothetical Interventions of Physical Activity on Obesity: Applying Target Trial Emulation in the 1970 Birth Cohort Study

Michail Katsoulis, Jamie Wong

Population Science & Experimental Medicine, Institute of Cardiovascular Science, UCL, London, UK

Background/Introduction Many studies examining the effect of physical activity on obesity have only used a single baseline measurement of physical activity in their analyses as their exposure. There is a limited number of papers that have evaluated the lifelong effect of physical activity on obesity measured at multiple timepoints, accounting for changes over time. This study aimed to use the target trial emulation framework to estimate the 16-year risk of obesity under hypothetical physical activity interventions in middle-aged individuals (30 to 46 years old) from the 1970 British Cohort Study (BCS70).

Methods This population-based cohort study utilised data from BCS70, a British birth cohort involving 8479 individuals born in 1970. Participants were allocated to one of two hypothetical interventions: low (0-3 times per month) and high (1-5 times per week) physical activity levels. Obesity was defined when body mass index was $30\text{kg}/\text{m}^2$. The study aimed to estimate the per-protocol effect, assessing the impact of physical activity if all participants had fully adhered to their assigned intervention. Multiple imputation was used to address missing data, creating 10 imputed datasets. Inverse probability weighting was used to account for adherence, using information from time-fixed and time-dependent confounders. Pooled logistic regression models were utilised to estimate the standardised risk curves. Non-parametric bootstrap with 200 samples for each imputed dataset was used to

estimate 95% confidence intervals, defined by the 2.5 and 97.5 percentiles of the pooled sample of the $10 \times 200 = 2000$ of risk estimates[1].

Results 2371 participants were assigned to the low physical activity (hypothetical) intervention, and 6108 to the high physical activity (hypothetical) intervention. The estimated standardised risk of obesity over 16 years was 30.6% (95% CI = 27.1%, 34.8%) for the low physical activity group, and 29.8% (95% CI = 28.3%, 31.2%) for the high physical activity group. The risk difference between the two hypothetical interventions after 16 years was -0.8% (95% CI = -5.0%, 2.9%)

Conclusion Individuals who adhered to a high level of physical activity exhibited a moderately lower risk of developing obesity between 30 to 46 years of age. This study was among very few cohort studies overall, to utilise the target trial emulation framework to evaluate hypothetical interventions.

Reference Schomaker M, Heumann C. Bootstrap inference when using multiple imputation. Stat Med. 2018 Jun;37(14):2252-2266

Design and Evaluation of Clinical Trials

Wednesday, 2025-08-27 09:00 - 10:30, Biozentrum U1.131

Chair: Babak Choodari-Oskooei

1: Clinical Trial Simulation: Planning, Implementation and Validation Principles

Kim May Lee¹, Babak Choodari-Oskooei², Michael J. Grayling³, Peter Jacko^{4,5}, Peter K. Kimani⁶, Aritra Mukherjee⁷, Philip Pallmann⁸, Tom Parke⁴, David S. Robertson⁹, Ziyang Wang¹⁰, Christina Yap¹¹, Thomas Jakl^{9,12}

¹King's College London (United Kingdom)

²MRC Clinical Trials Unit at UCL, University College London

³Johnson & Johnson

⁴Berry Consultants

⁵Lancaster University

⁶Warwick Medical School, University of Warwick

⁷Newcastle University

⁸Centre for Trials Research, Cardiff University

⁹MRC Biostatistics Unit, University of Cambridge

¹⁰University of Southampton

¹¹Institute of Cancer Research, University of London

¹²University of Regensburg

The adoption of complex innovative clinical trial designs has been increasing in recent years. These are trial designs that have one or more unconventional features, which aim to improve upon conventional trial designs. The motivation for these designs may not be difficult to follow, but their set-up and implementation is usually more challenging. Statistical properties of these designs can also be difficult to compute. Clinical trial simulation (CTS), which uses software to generate artificial data for learning, can be conducted to identify the (optimal) setting of a clinical trial, evaluate the design's statistical properties, perform "what-if" analyses, and compare different design set-ups and data analysis strategies, all of which contributes to a better understanding of the value of unconventional features before implementing the design in an actual clinical trial. It is also a tool for methodologists to investigate a novel design. Existing literature on simulation primarily focuses on the evaluation of statistical analysis methods, with less attention on the detailed specification and planning of CTS. This work presents a step-by-step planning process for CTS in the context of complex innovative trials. The target audience comprises both trial statisticians who are involved in designing and analysing clinical trials, and statistical methodologists who focus on the development of

trial designs and of analysis methods.

2: Early Phase Dose-Finding Designs for CAR-T Cell Therapies

Weishi Chen¹, Pavel Mozgunov¹, Xavier Paoletti²

¹University of Cambridge, United Kingdom

²Institut Curie, France

Background Chimeric Antigen Receptor (CAR)-T cell is an immunotherapy which revolutionised the treatment of relapsed/refractory lymphoma and leukemia. It is shown to have higher response rate, higher mid-to-long term overall survival, and lower toxicity than standard treatments. However, due to lack of dose-limiting toxicity (DLT) and unclear dose-effect relationship, traditional phase I designs of clinical trials cannot lead to accurate selections of the optimal dose (OD). Among the reported trials, the design of phase I clinical trials are frequently unclear, and early phase designs specifically for CAR-T cells are needed. Beside clinical outcomes, the CAR-T cell expansion from serial blood samples is measured at various timepoints. Two main profiles of cell-evolution have been reported: the injected cells exhaust and are progressively eliminated from the blood, or they proliferate and are maintained over some duration before being eliminated.

Methods We propose a novel early phase dose-finding design for CAR-T, using both toxicity and activity endpoints to locate the OD, the dose with highest activity among safe doses. The CAR-T cells expansion is used to indicate activity, which is more sensitive than traditional clinical responses. A Bi-Exponential model is used to model the expansion trajectory, which approximates the CAR-T cells growth dynamics, is simple enough to be estimated with small sample sizes, and is flexible enough to accommodate the reported cell-evolutions. Three criteria for activity are considered: 1) number of cells at specific time, 2) duration before all cells are eliminated, 3) area under the cell-expansion curve. A non-parametric benchmark has been developed to evaluate the performance of the proposed design.

Results Simulations show that the OD can be selected with high accuracy even under small sample sizes. All three activity criteria work well when the model is correctly specified. Furthermore, the model is robust under model-misspecification if the appropriate activity criterion is used. Sensitivity analyses show that the proposed design is robust against missing measurements of CAR-T cells expansion and increased noise level.

Conclusion Both toxicity and activity endpoints should be used for CAR-T cells, and the CAR-T cells expansion is a more sensitive and specialised measure for biological activity

compared to clinical outcomes. Depending on the activity criteria, higher doses do not necessarily give higher activities.

3: Fast Approximation of the Operating Characteristics in Clinical Trials

Susanna Gentile¹, Daniel Schwartz^{2,3}, Riddhiman Saha², Lorenzo Trippa^{2,3}

¹Department of Statistical Sciences, Sapienza University of Rome

²Department of Biostatistics, Harvard T.H. Chan School of Public Health

³Department of Data Science, Dana-Farber Cancer Institute

Motivation:

Evaluating operating characteristics (OCs), such as expected sample size, power, and type I error, is essential for designing clinical trials. These OCs guide critical design decisions and are vital for interactions between the study team, regulatory agencies, and other stakeholders.

Traditionally, OCs are estimated using Monte Carlo simulations. This approach is based on repeatedly sampling the clinical trial under a specific *scenario* selected by the researcher. The OCs are then computed as functions of the aggregated results. This approach, however, can be computationally expensive, especially for complex trial designs (e.g., adaptive trials) or models requiring intensive inference (e.g., MCMC-based Bayesian methods). These computational demands can make thorough design evaluation impractical.

Our proposal:

We introduce the *Q-approximation*, a method for rapidly approximating OCs by leveraging three key principles:

1. Many clinical trial designs adhere to the likelihood principle, meaning all necessary information for decision-making is contained in the likelihood function.
2. Under mild regularity conditions, the log-likelihood is approximately quadratic, and the likelihood is approximately Gaussian.
3. The distributions of the center and curvature of the quadratic approximation can be derived using standard asymptotic theory.

These considerations allow approximating OCs by directly simulating *likelihood functions*

instead of simulating entire datasets as in traditional Monte Carlo analyses. More specifically, the Q-approximation can be much faster than the Monte Carlo approximation for two main reasons:

- Dimensionality Reduction: Instead of simulating entire datasets, we sample two-dimensional vectors, representing the center and the curvature of the approximation.
- Fast computation of inferences: Because the approximate likelihood is Gaussian, inferences can be computed analytically (e.g., no need for MCMC in a Bayesian logistic model).

Applicability and results:

The Q-approximation can be applied to more complex settings, including multi-stage adaptive designs, Bayesian adaptive randomization, and trials incorporating external data. We demonstrate its effectiveness across three trial settings: (a) non-adaptive two-arm randomized controlled trials (RCTs), (b) adaptive RCTs leveraging external data, and (c) multi-arm RCTs with Bayesian adaptive randomization. Our results show that the Q-approximation provides accurate OC estimates and reduces computational time by up to a thousandfold compared to Monte Carlo methods. This speed-up can make it much more practical to explore operating characteristics thoroughly and recommend complex designs.

4: Quantifying the Effects of Screening – a Re-Randomisation Approach

Vichithranie Wasantha Madurasinghe¹, Bethany Shinkins¹, Keith R Abrams^{1,2}, Sian Taylor-phillips¹

¹Warwick Medical School, University of Warwick, United Kingdom

²Department of Statistics, University of Warwick, United Kingdom

Background

Due to the complex causal pathways involved, quantifying the effects of screening, particularly the effects of early detection, can be problematic. In the context of screening, modelling studies can be used as a way of linking intermediate and long-term health outcomes when RCT data are limited. A previously conducted methodological review of modelling studies, using intermediate trial outcomes for estimating morbidity/mortality reductions of screening

interventions, found serious methodological limitations which hinders their usability. Consequently, a new framework for quantifying the overall (i.e. combined effects of early detection and treatment) and early detection effects of screening is introduced.

Methods

In a typical screening trial participants allocated to screening are invited for testing while control participants are tested when they present with symptoms. In both groups those who are detected are treated using similar treatment protocols. Therefore, if the patients referred for treatment are prognostically similar between groups, there is no reason to expect the treatment effects to differ between control and screen detected patients.

If trial participants are randomised for a second time, then an unbiased effect estimate of early detection can be derived by comparing the number of events observed in screened arm as per original randomisation to expected number of events in re-randomised control arm. The overall effect of screening can be estimated by comparing the number of outcome events in screen compared to control groups as per original randomisation.

Such a re-randomisation approach to estimating the effect of screening can be implemented using simulation. The feasibility of such an approach is illustrated using data extracted from five cancer (breast, colorectal, cervical, lung, and prostate) screening trial publications.

Results

The expected mortality reductions, estimated using a re-randomisation approach using intermediate outcomes, are in-line with mortality reductions observed in the trials with the ratio of relative risks (observed to expected) ranging from 0.66 to 0.99. However, there was a greater variability in the estimated expected mortality reductions when there was a larger difference in number of cases with late-stage disease (i.e. lymph node and/or distant metastatic disease at the first presentation) between the trial arms.

Conclusion

The framework introduced and applied here provides a unified analysis approach for quantifying the overall and early detection effects of screening. A simulation study to assess the performance of the proposed approach across a range of scenarios is on-going.

5: From Methodology to Mindset: Implementing Quantitative Decision-Making Framework into Early Development Clinical Trials

Stefan Englert¹, Leen Slaets², Lilla Di Scala³

¹ Janssen-Cilag GmbH, a Johnson & Johnson company, Germany

² Janssen Pharmaceutica NV

³ Actelion Pharmaceuticals Ltd, a Johnson & Johnson Company

Background Early-phase clinical trials in oncology are characterized by high uncertainty and small sample sizes, making decision-making challenging. The overall success rate of development programs, as measured by the likelihood of approval, was previously estimated to be ~26% (DiMasi et al 2005). Informed decision making is therefore crucial for success, both in terms of speed and quality of the decisions taken. Focus has shifted from traditional hypothesis testing to quantitative decision-making frameworks that acknowledge the early development challenges. Such frameworks not only enhance the decision-making process but also empower statisticians to contribute to critical development decisions.

Methods A Quantitative Decision-Making (QDM) framework was established which leverages Bayesian approaches to quantify the probability of success in early-phase studies. It integrates relevant criteria and risk thresholds based on internal targets (target-product-profile), as well as external/published data. The goal of the framework is to offer statistical guidance for cohort expansion or discontinuation, and possible transition to late development.

The talk will detail the steps taken to integrate methodology into clinical planning practices and to expand the role of statisticians.

Results The implementation of the QDM framework facilitates a systematic evaluation of (preliminary) efficacy based on empirical evidence obtained in the ongoing early-phase clinical trial(s) as well as from external or prior evidence, when available. Based on these data the framework assesses the probability of achieving clinically meaningful outcomes (DiScala et al 2013). Using pre-defined criteria and thresholds, decisions regarding continuation or stopping studies are better informed by quantification of the risks involved.

The talk will demonstrate the calculation of decision criteria for a Phase I/II study and show how the operating characteristics are assessed to ensure the robust decision criteria. The framework has been implemented within governance meetings, providing a pathway for statisticians to actively shape the clinical development strategy.

Conclusions The QDM framework enhances the decision-making process in early development oncology by providing a robust and actionable analytical approach to quantify uncer-

Abstracts of Contributed Talks

tainty and risk. By adopting this framework, drug development can become more consistent, allowing for informed decisions that optimize the potential for successful clinical outcomes in a competitive landscape.

The talk will address challenges to widespread adoption, projecting increased use of such decision criteria in the future.

References DiMasi, Grabowski. Economics of new oncology drug development. *J Clin Oncol*, 25(2007)

Di Scala, Kerman, Neuenschwander. Collection, synthesis, and interpretation of evidence: a proof-of-concept study in COPD. *Statistics in Medicine*, 32(2013)

Competing Events and Multi-State Modelling

Wednesday, 2025-08-27 09:00 - 10:30, Biozentrum U1.141

Chair: Sarah Friedrich

1: A General Approach to Fitting Multistate Cure Models Based on an Extended-Long-Format Data Structure

Yilin Jiang^{1,2}, Harm van Tinteren², Marta Fiocco^{1,2,3}

¹Leiden University, the Netherlands

²Princess Maxima Center, the Netherlands

³Leiden University Medical Center, the Netherlands

A multistate cure model is a statistical framework used to analyze and represent the transitions individuals undergo between different states over time, accounting for the possibility of being cured by initial treatment. This model is particularly useful in pediatric oncology where a proportion of the patient population achieves cure through treatment and therefore will never experience certain events. Traditional multistate models do not account for this population heterogeneity. A multistate cure model can provide a more comprehensive understanding of the disease progression or remission patterns and aid in personalized treatment decisions and prognosis prediction. Despite its importance, no universal consensus exists on the structure of multistate cure models. Our study provides a novel framework for defining such models through a set of non-cure states. We develop a generalized algorithm based on the extended long data format, an extension of the traditional long data format, where a transition can be divided into two rows, each with a weight assigned reflecting the posterior probability of its cure status. The multistate cure model is built upon the current framework of multistate model and mixture cure model. The proposed algorithm makes use of the Expectation-Maximization (EM) algorithm and weighted likelihood representation such that it is highly flexible in model specification and easy to implement with standard packages. Additionally, it facilitates dynamic prediction. The state occupancy probabilities can be easily obtained within our multistate cure model framework, incorporating baseline covariates or even post-baseline information. The algorithm is applied on data from the European Society for Blood and Marrow Transplantation (EBMT). Standard errors of the estimated parameters in the EM algorithm are obtained via a non-parametric bootstrap procedure, while the method involving the calculation of the second-derivative matrix of the observed log-likelihood is also presented.

2: Calibration of Cause-Specific Absolute Risk for External Validation using Each Cause-Specific Hazards Model in the Presence of Competing Events

Sarwar Islam Mozumder^{1,2}, Sarah Booth¹, Richard Riley³, Mark Rutherford¹, Paul Lambert^{4,5}

¹Biostatistics Research Group, Department of Population Health Sciences, University of Leicester, Leicester, United Kingdom

²AstraZeneca UK, Cambridge, United Kingdom

³Institute of Applied Health Research, College of Medical and Dental Sciences, University of Birmingham, Birmingham, United Kingdom

⁴Cancer Registry of Norway, Norwegian Institute of Public Health, Oslo, Norway

⁵Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

Background Calibration is assessed on cause-specific absolute risks to determine agreement between predicted risks from the model and observed risks. For competing risks data, correct specification of more than one model may be required to ensure well-calibrated predicted risks for the event of interest. Furthermore, interest may be in the predicted risks of the event of interest, competing events and all-causes. Therefore, calibration must be assessed simultaneously using various measures.

Objectives Evaluate calibration of prediction models for external validation using the cause-specific hazards (CSH) approach

Methods We propose assessing miscalibration for CSH models using the complement of the cause-specific survival alongside assessment of calibration of the cause-specific absolute risks. We simulated various scenarios to illustrate how to identify which model(s) is mis-specified in an external validation setting. Calibration plots and calibration statistics (Calibration Slope, calibration-in-the-large) are presented alongside performance measures such as the Brier Score and Index of Prediction Accuracy. We propose using pseudo-values to calculate observed risks and we generate a smooth calibration curve with restricted cubic splines. We fitted flexible parametric survival models to the simulated data to flexibly estimate baseline CSH for prediction of individual cause-specific absolute risk. Methods will also be illustrated using a clinical dataset.

Results Our simulations illustrate that miscalibration due to changes in the baseline CSH in external validation data are better identified using components from each cause-specific model. A mis-calibrated model on one cause, could lead to poor calibration on predicted

absolute risks for each cause of interest, including the all-cause absolute risk. This is because prediction of a single cause-specific absolute risk is impacted by effects of variables on the cause of interest and competing events.

Conclusions If accurate predictions for both all-cause and each cause-specific absolute risks are of interest, this is best achieved by developing and validating models via the CSH approach. For each cause-specific model, researchers should evaluate calibration plots separately using probabilities obtained using each respective cause-specific model components to reveal the cause of any miscalibration. Pseudo-values are also proposed to obtain observed individual cause-specific net or absolute risk and smoothed calibration curves.

3: Multi-State Models with Restricted Transition Windows: The Impact of Time Scale Choice

Georgy Gomon^{1,2}, Ilaria Prosepe¹, Rachel Knevel^{2,3}, Saskia le Cessie^{1,4}

¹Department of Biomedical Data Sciences, Leiden University Medical Centre, Leiden, the Netherlands

²Department of Rheumatology, Leiden University Medical Centre, Leiden, the Netherlands

³Rheumatology, Newcastle University Translational and Clinical Research Institute, Newcastle upon Tyne, UK

⁴Department of Clinical Epidemiology, Leiden University Medical Centre, Leiden, the Netherlands

Introduction

Multi-state models describe processes where individuals transition between states over time. An important modelling choice is the time-scale, with two common choices available: **(i)** "clock-forward", measuring time from initial state entry, or **(ii)** "clock-reset", resetting time at each state entry.

Transition probabilities often depend on time of state entry. The "clock-forward" method incorporates this through delayed entry but struggles when some transitions are restricted to specific time windows after state entry. Such restricted transitions, where events occur only within a specific time-window after state entry, are common in medical settings, such as referrals at consultations, treatment initiation shortly after diagnosis, or early medication side effects. The "clock-reset" approach captures how the transition intensities of restricted transitions are equal to zero after the restricted time period has passed, but necessitates incorporating time of state entry as time-dependent covariate, complicating prediction.

Methods We consider a multi-state model where time since initial entry affects transition probabilities, and some transitions occur only within restricted time windows. Our motivating example is the referral of patients with musculoskeletal complaints from general practitioners (GP) to specialists. Patients may have up to five consultations for a complaint, with referrals possible only during consultations. Between consultations, referral probability is zero.

We compared two time-scale approaches:

1. **Clock-Forward:** Time is measured from the initial consultation.
2. **Clock-Reset:** Time resets at each consultation, with time since first consultation included as time-dependent covariate. This approach captures how the baseline hazard of time-restricted transitions (referral) is equal to zero post-consultation.

Both models included baseline and time-dependent covariates, with the GP consultations as intermediate states and referral as absorbing state. Event probabilities were estimated by generating paths through the model. Model performance was assessed using discrimination (c-index) and calibration (O/E ratio & calibration plot).

Results We analyzed 2,358,750 primary care episodes, with 1,842,533 transitions and 23,884 referrals. The choice of time-scale significantly influenced estimated covariate effects for time-restricted transitions and predicted event probabilities. A simulation study will be presented that further investigates strategies for incorporating time-restricted transitions in multi-state models.

Conclusion We present insights into how the time-scale selection (“clock-forward” or “clock-reset”) impacts multi-state models, particularly for transitions with restricted time windows. In our application, the two approaches yield different results in terms of estimated covariate effects and predicted probabilities. The simulation study will give more insight into modeling strategies for time-restricted transitions in multi-state model.

4: Integrating Landmarking and Competing Risks in Survival Analysis with Machine Learning Techniques

Shirin Sultana, Dr. Md Hasinur Rahaman Khan

Institute of Statistical Research and Training (ISRT), University of 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 Results show that in terms of calibration, discrimination, and prediction accuracy, mixed-model landmarking routinely performs better than standard landmarking. Simulation studies highlight the necessity for specialized models by showing that ignoring competing hazards significantly overestimates occurrence probabilities. Mixed-model landmarking outperforms RSF in terms of prediction, achieving lower Brier scores (as low as 0.021) and higher AUC values (up to 0.883). These results are supported by real-world data, which demonstrates how well longitudinal data and machine learning can be combined for dynamic prediction. RSF stands out as the most reliable machine learning technique, outperforming gradient-boosting models in terms of calibration and discrimination.

Conclusion This study illustrates the benefits of mixed-model landmarking within competing risk frameworks, filling a significant gap in survival analysis. Predictive accuracy is improved by integrating cutting-edge machine learning techniques, providing reliable solutions for a range of clinical applications. Besides, dynamic modeling frameworks driven by machine learning are useful tools for enhancing survival prediction in medical research because they can handle high-dimensional data and nonlinear interactions with ease.

5: Survival without vs after Transition to the Intermediate Event in a Non-Markovian “illness-Death” Model: Application to Heart Transplant Data

Davide Paolo Bernasconi^{1,2}, Lorenzo Del Castello¹, Laura Antolini¹

¹University of Milano-Bicocca, Italy

²ASST Grande Ospedale Metropolitano Niguarda, Italy

Introduction The “illness-death model” describes a simple multistate process where subjects can move from the initial state to the final state (death) possibly transiting to an intermediate state (illness). This framework applies also when the intermediate state consists in a therapeutic intervention administered after waiting some time since the initial event, e.g. cardiopathic patients waiting for a possible heart transplant.

Methods When the aim is to estimate and compare the survival of patients on the waiting list vs after being transplanted, it is important to check the validity of the Markov assumption in order to select the most appropriate time scale (clock-forward or clock-reset) guiding the mortality after transplant and to assess the role of the waiting time [1].

We analyzed data from a cohort of nearly 1000 patients affected by severe cardiomyopathy included in a waiting list for heart transplant but with low priority. The process is non-markovian because the mortality rate since transplant tends to increase along with a longer waiting time.

Results We extended a non-parametric method to estimate survival in the presence of a time-dependent intervention [2], accounting for the impact of waiting time, to answer the question: what is the survival of a patient that will never be transplanted compared to a patient transplanted immediately after entry in list? Using a landmark approach, we also answer questions like: what is the survival of a patient alive at a certain time after entry on list and that will never be transplanted compared to a patient alive and transplanted at that time?

Finally, we used predictions from a double-scale Cox model [3] to estimate profile-specific survival without vs after transplant, adjusting for baseline (i.e. at entry on list) covariates.

Conclusion We show an original approach for survival prediction in the presence of a time-dependent treatment covariate accounting for the waiting time until treatment switch in a non-markovian process.

References [1] Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine* 2007; 26:2389-2430.

[2] Bernasconi DP, Rebora P, Iacobelli S, Valsecchi MG, Antolini L. Survival probabilities with time dependent treatment indicator: quantities and non-parametric estimators. *Statistics in Medicine* 2015; 35(7):1032-48.

[3] Tassistro E, Bernasconi DP, Rebora P, Valsecchi MG, Antolini L. Modelling the hazard of transition into the absorbing state in the illness-death model. *Biometrical Journal* 2019;

Abstracts of Contributed Talks

62(3):836-851.

Machine Learning 2

Wednesday, 2025-08-27 09:00 - 10:30, Biozentrum U1.101

Chair: Wessel van Wieringen

1: Asymmetric Shapley Values to Quantify the Importance of Genomic Variables in Clinical Prediction Studies

Mark van de Wiel

Amsterdam UMC, Netherlands, The

In many clinical prediction studies, genomics data are combined with other variables, such as gender, sex and disease state. In such models, the additional predictive strength of genes is often disappointingly small. This may lead to the premature conclusion that genomics is irrelevant for the disease or its progression. Quantifying the 'importance' of genomic variables purely by assessing decreased prediction accuracy after removal from these variables to the model has two shortcomings. First, it does not account well for correlations. If those genomic variables are highly correlated to the other (low-dimensional) variables, the latter act as a buffer when removing genomics from the model, leading to little decline in performance. Still, the two types of variables may be equally important for the outcome. Shapley values, either for single variables or groups of variables, provide a solution for this in the sense that importance is spread out over correlated variables. A second shortcoming is that causality or temporal ordering is completely ignored. This is particularly relevant, as genomic aberrations are often conceived to be close to the root cause. This means that they may have indirect effects on the outcome via some of the other variables.

Motivated by an application to prediction of relapse-free survival for colorectal cancer patients, we study asymmetric Shapley values to quantify the importance of genomic variables when disease state acts as a mediator, and when several confounders are present. We show that accounting for the ordering matters for quantifying the importance of genomics. Moreover, we address regularization to deal with dependency of the mediator (and confounders) on the high-dimensional genomics variable. These dependencies are crucial for the calculation of Shapley values. Finally, we illustrate how to perform inference with the asymmetric Shapley values to compare the importance of genomics with that of other variables. Extensions to pre-defined gene sets will be also be discussed. To summarize, this work helps researchers to obtain a better quantification of the (relative) importance of genomics variables in clinical prediction settings.

2: Fused Estimation of Varying Omics Effects for Clinico-Genomic Data

Jeroen Goedhart¹, Mark van de Wiel¹, Wessel van Wieringen^{1,2}, Thomas Klausch¹

¹Amsterdam University Medical Centers, Netherlands, The

²Vrije Universiteit Amsterdam, Netherlands, The

Background

Cancer prognosis is often based on a set of omics covariates and a set of established clinical risk factors such as age, tumor stage, and prognostic indices. Combining these two sets of covariates in a so-called clinico-genomic model poses challenges. First, difference in dimension: clinical covariates should be favored because they are low-dimensional and usually have stronger prognostic ability compared to high-dimensional omics covariates. Second, complex interactions: since many cancers are heterogeneous, genetic profiles and their prognostic effects may vary across patient subpopulations. Last, redundancy: a (set of) gene(s) may encode similar prognostic information as a classical risk factor.

Methods

To address these challenges, we combine regression trees, employing clinical covariates only, with a unique fusion-like penalized regression framework in the leaf nodes for the omics covariates. The fusion penalty controls the modeled variability in genetic profiles across subpopulations defined by the tree. We prove that the shrinkage limit of our penalized framework equals a benchmark model for clinico-genomic data: a ridge regression with penalized omics covariates and unpenalized clinical risk factors. Along with boosting prognostic performance for various situations, the proposed method has another practical advantage. It allows researchers to evaluate, for different patient subpopulations, whether the added overall omics effect enhances prognosis compared to only employing clinical covariates.

Results and Conclusion

We illustrate the strengths of the proposed method in simulations and in an application to colorectal cancer prognosis based on age, tumor stage, gender, a molecular clustering variable, and 20,000+ gene expression measurements. Our method reveals that the overall omics effect is not required for colorectal cancer prognosis of some patient subpopulations. Our method also finds a large variability in the subpopulation-specific effects of a set of genes related to expression of cancer/testis antigens.

3: Random Forests using Longitudinal Predictors

Justine Remiat¹, Cécile Proust-Lima², Robin Genuer¹

¹Univ. Bordeaux, INSERM, INRIA, BPH, U1219, France

²Univ. Bordeaux, INSERM, BPH, U1219, France,

Introduction and Objectives Random Forests are an effective predictive tool, particularly in high-dimensional settings. However, they are not well-suited for longitudinal data collected over time. To address this limitation, Fréchet Random Forests [1] were proposed. They can handle any type of data within a metric space by using a distance tailored to each data type (e.g., images, trajectories). This work aimed to implement the Fréchet Random Forest for trajectory data, fully exploiting the flexibility of the Generalized Fréchet distance; and evaluate the performance of the Fréchet Random Forest in predicting a continuous outcome using longitudinal inputs.

Methods The generalized discrete Fréchet distance depends on a time-shifting parameter, called timescale, which modifies its behavior. We proposed two implementations: the timescale defined as an hyper parameter or the time-scale randomly drawn at each tree node to explore all time sensitivity behaviors. A simulation study has been conducted to illustrate the flexibility of the Fréchet random forest to capture different scenarios of association:(i) time-sensitive association (ii) shape-sensitive association and (iii) a mix of both. We then apply the method to data from a population-based cohort to predict the risk of dementia from clinical marker trajectories.

Results The simulations illustrated the flexibility of the Fréchet Random Forests to adapt to different types of associations with the timescale tuning. The Fréchet forests also demonstrated better predictive performance (MSE) across all three scenarios compared to classical Random Forests with pre-determined features. On the application data, the Fréchet forests outperformed classical forests, even with more irregular and sparse data, while similarly identifying predictive markers.⁶

Conclusion Thanks to its tunable timescale parameter that can adapt to different structures of association, the Fréchet Random Forest constitutes a flexible tool for prediction based on longitudinal data.

[1] Capitaine L. et al. Fréchet Random Forests for Metric Space Valued Regression with Non-Euclidean Predictors. JMLR. 2024

4: Digital Twins you can ‘count’ On: a Novel Application of Digital-Twin Prognostic Scores in Negative-Binomial Models.

Tasos Papanikos, Doug Thompson, Harry Parr, Aris Perperoglou

GlaxosmithKline, United Kingdom

Introduction Methods for including prognostic information in clinical trials have seen a resurgence via the development of so-called ‘digital-twins’ (DT) – whereby an ensemble of machine-learning models are trained on historical data before being then adopted within the analysis of a current trial. The potential of controlling statistical uncertainty while preserving asymptotic unbiasedness of the marginal treatment estimator could be positively seen by regulators. Still, applications in non-linear outcomes such as binary, count or time-to-even endpoints require more understanding and addressing issues around non-collapsibility. More recently Conner *et al.* [1] evaluated the collapsibility property of the Rate Ratio (RR) treatment effect, concluding that the marginal RR is equal to the conditional RR when covariates are prognostic (i.e., with no heterogeneity of the treatment effect). The aim of this work is to evaluate the use of DTs as a novel application in the analysis of count data, and its utility in trials for respiratory diseases such as COPD, where the primary analysis is frequently a Negative-Binomial or Poisson regression model.

Methods We evaluate the performance of DT models for count or recurrent endpoints by carrying out an extensive simulation study. We take inspiration from COPD trial data to motivate a diverse set of plausible scenarios. We consider varying the design parameters of the present clinical trial, whilst additionally exploring how a realistic drift in performance (i.e., historical data versus current trial data) may impact key operating characteristics (e.g., power, type 1 error, bias etc.).

Results A full set of results will be presented to illustrate the expected effect of applying DT on sample size, and what might be realistic expectations under each simulation scenario. Additionally, we will present whether the expected consistency between the conditional and marginal treatment effect estimates holds and where further investigations may be expected.

Conclusion Through this work we have demonstrated a novel application of DT methodologies for count or recurrent endpoints. Our work highlights new areas of expansion for DTs and indicates possible ‘quick wins’ for leveraging prognostic information in COPD trials.

5: Calibrating Machine Learning Approaches for Probability Estimation in Case of the Absence of Calibration Data

Eleonora Di Carluccio¹, Giorgos Koliopanos¹, Francisco Ojeda^{3,4}, Christian Weimar², Andreas Ziegler^{1,3,4,5}

¹Cardio-CARE, Switzerland

²BDH-Klinik Elzach, Elzach, Germany

³Department of Cardiology, University Heart and Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

⁴Centre for Population Health Innovation (POINT), University Heart and Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

⁵School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Pietermaritzburg, South Africa

Statistical prediction models are gaining popularity in applied research. One challenge is the transfer of the prediction model to a population that may be structurally different from the one for which the model was developed. If a cohort is available from the target population, the model can be calibrated to the characteristics of the target population. Elkan proposed a closed formula for calibration in case a calibration cohort is lacking. His method relies on the equal distribution of covariates in affected as well as unaffected individuals for both the existing and the calibration cohort. In this presentation, we propose a novel method which uses synthetic data generation from an existing cohort in conjunction with marginal statistics from the calibration cohort. A “recalibration” logistic model is computed in the synthetic data and used to recalibrate the predicted probabilities in the calibration cohort. We illustrate the novel approach in a simulation study and with two real data sets. The simulation studies and the illustration demonstrate the potential of this novel approach for calibration in absence of calibration data, when marginals are correctly specified and the correlations between variables from the model development and the population for which calibration is required are identical.

Prediction / Prognostic Modelling 3

Wednesday, 2025-08-27 09:00 - 10:30, ETH E27

Chair: Carolin Strobl

1: Predicting Prediction Performance

Max Westphal, Rieke Alpers

Fraunhofer Institute for Digital Medicine MEVIS, Germany

Introduction Internal-external validation studies aim to quantify the transferability of a clinical prediction model to a new context or population. In a recent work, we introduced an estimand framework to allow a precise specification of the difference between development and inference context and in effect the exact type of transferability to be estimated. [1] The chosen estimand has direct implications for the data splitting scheme in the validation study. In this talk, we will focus on comparing different statistical models for the out-of-distribution performance of a newly developed prediction model or learning algorithm. Such models can be used for a “meta” prediction of the unknown performance of the new “primary” prediction model or algorithm once it is implemented in a new context (e.g., in a country that was not part of the development data) and in particular provide an adequate uncertainty quantification for this prediction.

Methods In a case study based on the International Stroke Trial dataset, we compared different statistical modelling approaches to predict (out-of-distribution) prediction performance. Hereby, performance was measured as the area under the curve (AUC) of the developed classification models for two-week survival after an acute stroke. [2] We compared a frequentist meta-analysis and a variety of Bayesian hierarchical models. For the latter, different (linear and non-linear) parametric learning curve models were utilized to model the dependence of the performance on the training sample size.

Results and Discussion The frequentist meta-analysis approach is relatively simple to apply and provides plausible performance predictions. However, in contrast to the Bayesian hierarchical modelling approach, it cannot be used directly to adequately characterize the dependence on the training sample size. It should thus only be used if training sample sizes in the development dataset are representative. Another advantage of the Bayesian approach is that posterior uncertainty estimates offer a more expressive description of the different levels of uncertainty in the data (e.g., unseen patients, clinics or countries). It is however

also more complex to use, requiring the specification of the hierarchical model structure, prior distributions and a learning curve model.

References [1] Alpers, Rieke, and Westphal, Max. (2025). “An estimand framework to guide model and algorithm validation in predictive modelling.” Submitted for publication.

[2] International Stroke Trial Collaborative Group. (1997). The International Stroke Trial (IST): a randomised trial of aspirin, subcutaneous heparin, both, or neither among 19 435 patients with acute ischaemic stroke. *The Lancet*, 349(9065), 1569-1581.

2: Predictive Modelling with Block Missingness

Viktor Racinskij¹, Robin Mitra², Chris Harbron³, Niels Hagenbuch⁴

¹The Alan Turing Institute (United Kingdom)

²UCL (United Kingdom), The Alan Turing Institute (United Kingdom)

³Roche Pharmaceutical (United Kingdom)

⁴F. Hoffmann-La Roche AG (Switzerland)

There is increasing interest in creating and utilising large data sets combined at scale across different modalities. Sophisticated predictive modelling methods can routinely handle large numbers of possible predictors derived from such data sources. However, when some units are not measured on certain modalities, missing data blocks will likely arise in the integrated data and can be viewed as an instance of structured missingness. These complicate the application of predictive models on such data sources.

Multiple imputation is a convenient and theoretically justified tool to handle missing values in many settings, including predictive modelling. However, in highly multivariate settings with large blocks of missing data, the approach may be computationally demanding, and it is increasingly difficult to specify a well-performing imputation model. In contrast, we leverage the basic principle of the factored likelihood approach and combine the joint model of missing data given the observed with the marginal statistics related to the observed data. Our method directly adjusts for missingness in statistics that play a role in predicting the outcome variable. In doing so, we can make predictions based on all the observed data but without the need to perform any imputation and without explicit specification of the likelihood function.

We compare our factorisation-based approach with multiple imputation (such as drawing from a predictive distribution) and regression imputation (using the conditional mean) across

a range of comprehensive simulations. Our method avoids bias incurred by a single regression-type imputation, and its achieved predictive accuracy matches or exceeds that of carefully implemented multiple imputation, while computational burden is substantially reduced. We also illustrate our approach on a set of variables from the Cancer Genome Atlas, a large-scale multimodal pan-cancer data set, and identify some key insights from this application.

The benefits of the proposed method include its computational and statistical efficiency, with no need to create multiple imputations to obtain unbiased predictions. It also does not rely on distributional assumptions about the underlying data. In conclusion, our method provides a fast, reliable, and accurate solution for predictive modelling in cases of block missingness.

3: Lifestyle Predictors of All-Cause Mortality: Enhancing Risk Models Beyond Traditional Biomarkers

Yuhe Wang^{1,2}, Cameron Razieh^{1,2}, Thomas Yates^{1,2}

¹Diabetes Research Centre, University of Leicester, Leicester General Hospital, Leicester LE5 4PW, UK

²NIHR Leicester Biomedical Research Centre, University of Leicester and University Hospitals of Leicester NHS Trust, LE5 4PW, UK

Background Accurate risk prediction models are crucial for estimating all-cause mortality, particularly among aging populations with a high prevalence of chronic diseases. Traditional models primarily rely on non-modifiable factors and clinical biomarkers such as blood pressure and cholesterol-to-HDL ratio. However, lifestyle factors including physical activity, strength, and fitness measures may provide additional predictive value or serve as alternatives in the risk prediction model. This study aims to assess whether substituting traditional biomarkers with easily measurable lifestyle factors can improve mortality risk prediction.

Methods Data were obtained from the UK Biobank and stratified by disease history, sex, and age (using 60 years as the threshold). The base model included six traditional risk factors: age (years), smoking status (Never, Previous, Current), BMI (kg/m^2), systolic blood pressure (BP) (mmHg), total cholesterol-to-HDL ratio (mmol/L), and Townsend deprivation score. Five lifestyle factors include resting heart rate (RHR), handgrip strength (HGS), leisure-time physical activity (LTPA), walking pace (WP), and sleep duration were incorporated either as additions to or replacements for traditional risk factors to the base model, individually or combined. The analysis was conducted in three stages: (1) adding lifestyle factors, (2) substituting BP or cholesterol-to-HDL ratio, and (3) replacing both BP and cholesterol-to-HDL ratio. Model performance was evaluated using the C-index, comparing lifestyle risk

predictors incorporated models to the traditional clinical model.

Results Adding lifestyle factors and substituting the cholesterol-to-HDL ratio improved risk discrimination across all subgroups, with the greatest improvements observed when incorporating all five lifestyle predictors together. In the unhealthy cohort, the C-index improved by 0.0360 in young women, 0.0161 in older women, 0.0227 in young men, and 0.0237 in older men when replacing cholesterol-to-HDL ratio with lifestyle factors. A similar pattern was observed in the healthy group, though with smaller differences between the base and cholesterol-substituted models. Overall, RHR provided the greatest predictive improvement, except in healthy women, where HGS showed the highest predictive enhancement.

Conclusion Lifestyle-based predictors, particularly RHR and combined 5 lifestyle factors, enhance mortality risk prediction and may serve as viable alternatives to clinical biomarkers. Given their accessibility and non-invasive nature, these factors could be integrated into prognostic models to improve risk estimation, particularly in settings with limited clinical data. Further research is needed to confirm the long-term utility of lifestyle predictors in mortality risk assessment.

4: Development and Validation of the Options Model, a Clinical Prediction Model Predicting Risk of Emergency Caesarean Births in Nulliparous Women

Alexandra Hunt

University of Liverpool, United Kingdom

Objective Globally, the rate of caesarean births (CB), including emergency caesarean births (EmCB), is increasing significantly. It is estimated that nearly one-third of all births will involve caesareans by 2030.

Several tools exist to predict Emergency Caesarean Births (EmCB), but they are not yet routinely implemented in clinical practice. While these tools generally demonstrate acceptable performance, external validation of some models and changes to national guidelines highlight the need for a new prediction model applicable to all-risk women. Using routinely collected data from a multi-ethnic pregnancy cohort in Bristol, the Options study aimed to develop and externally validate a clinical prediction model predicting the risk of EmCBs in nulliparous women and introduced a point scoring system to make the model more accessible and easily understood. This innovation promotes a more personalised and relaxed discussion between expecting mothers and their midwives. This model is validated across three diverse UK populations to ensure broad applicability.

Methods The model includes predictors age, height, BMI, estimated fetal weight, and weight gain. The model was developed using multivariable fractional polynomials and data funded by NIHR Bristol BRC, encompassing approximately 26,600 records from pregnant women at NBT in Bristol since 2009. External validation was conducted using datasets from Born in Bradford, Cambridge and Liverpool. Discrimination and predictive performance were assessed through C-statistics and calibration plots. A trial phase of the study will be implemented via a point scoring system, utilising the tool within a clinical setting.

Results The Options model demonstrates good internal discriminative ability (C-statistic: 0.66) and strong calibration. External validation results are comparable, showcasing good generalisability across diverse UK populations.

Conclusions Ensuring the applicability of prediction models across heterogeneous populations is essential. The Options prediction model forecasts EmCBs at 36 weeks gestation, supporting evidence-based, personalised care for pregnant women across varying risk profiles, but underlying differences in prevalence across cohorts highlight the challenges posed by varying regulations and hospital preferences. The prediction model has the potential to be integrated into NHS clinical practice, as a point scoring system, facilitating informed discussions between women and their clinicians regarding labour planning at 36 weeks.

5: Optimizing Dynamic Predictions from Joint Models using Super Learning

Dimitris Rizopoulos¹, Jeremy M.G. Taylor²

¹Erasmus MC, the Netherlands

²University of Michigan, USA

Background The motivation for our research comes from prostate cancer patients who, after diagnosis, underwent surgical removal of the prostate gland. The treating physicians closely monitor the prostate-specific antigen (PSA) levels of these patients to determine the risk of recurrence and metastasis and determine reintervention. Joint models for longitudinal and time-to-event data have been previously employed in prostate cancer to calculate dynamic individualized predictions and guide physicians. Two components of joint models that influence the accuracy of these predictions are the shape of the longitudinal trajectories and the functional form linking the longitudinal outcome history to the hazard of the event.

Methods Finding a single well-specified joint model that produces accurate predictions for all subjects and follow-up times can be challenging, especially when considering multiple longitudinal outcomes. In this work, we use the concept of super learning and avoid selecting

a single model. In particular, we specify a weighted combination of the dynamic predictions calculated from a library of joint models with different specifications. In particular, we focus on various formulations of the time effect for the longitudinal outcome and different functional forms to link this outcome with the event process. The weights are selected to optimize a predictive accuracy metric using V-fold cross-validation. We use as predictive accuracy measures the expected quadratic prediction error and the expected predictive cross-entropy.

Results In our motivating University of Michigan Prostatectomy Dataset, the ensemble super learner performed better than the model best-selected model in the cross-validation procedure, especially when using the expected predictive cross-entropy as an accuracy metric. In a simulation study, we found that the super learning approach produces results very similar to those of the Oracle model, which was the model with the best performance in the test datasets. All proposed methodology is implemented in a freely available R package.

Infectious Disease and Longitudinal Modelling

Wednesday, 2025-08-27 09:00 - 10:30, ETH E23

Chair: Hein Putter

1: Flexible Parametric Additive Hazards Regression for Modeling Excess Mortality in Pandemics

Liesbeth C de Wreede¹, Marina T Dietrich², Ilaria Prosepe¹, Hein Putter¹, Mar Rodriguez-Girondo¹

¹LUMC (the Netherlands)

²University of Augsburg (Germany)

Effective decision-making during a pandemic requires balancing hospital capacity, protecting vulnerable groups, and minimizing societal burden. To make optimal policy decisions, it is crucial to quantify the impact of the pandemic on (excess) mortality and to assess the effects of public health interventions across various groups, accounting for infection and vaccination dynamics. However, a major challenge in the analysis of pandemic data is the absence of reliable cause-of-death information, which complicates the assessment of the pandemic's specific impact.

Relative survival techniques are often used to assess excess mortality in a specific patient population by splitting observed mortality into background and excess mortality. These methods have been widely used to estimate cancer-specific mortality without the need for precise cause-of-death data. However, applying these techniques to estimate excess mortality in pandemic settings presents special challenges: standard relative survival methods assume (1) excess hazards are always positive, (2) background mortality is specified externally rather than derived from the data and based solely on demographic factors, and (3) a single time scale suffices. Pandemics, however, can involve negative excess hazards due to the protective effects of public health measures, large variations in background mortality across groups, and multiple time scales, such as time since vaccination and infection, that impact mortality risk beyond the main calendar time scale. Moreover, unlike in settings where the study population is a small subset of the reference population, in pandemic settings the background and study populations refer to the same group observed at different time points (pre-pandemic vs. pandemic).

To address these challenges, we propose a novel flexible parametric additive hazards model for pre-pandemic and pandemic data, using B-splines to estimate baseline hazards and time-

dependent covariate effects. This new perspective on excess mortality estimation through a single additive hazards model accommodates negative excess hazards and integrates relevant risk factors into background mortality, equal to pre-pandemic mortality and estimated from the data. Moreover, this approach can handle multiple time scales and is less prone to overfitting than the classical non-parametric Aalen's method. We investigated two estimation procedures: one based on the least-squares approach used for Aalen's method, and the other exploiting the equivalence between additive hazards models with piecewise constant hazards and Poisson models with an identity link. The performance of the approach is demonstrated through a simulation study based on the COVID-19 pandemic and scenarios mimicking plausible pandemic conditions.

2: Investigating the Association Between Risky Sexual Behaviors and HIV Risk Using Multivariate Joint Models

Nobuhle Nokubonga Mchunu¹, Henry Mwambi², Tarylee Reddy¹, Nonhlanhla Yende-Zuma¹, Dimitris Rizopoulos³

¹Biostatistics Research Unit, South African Medical Research Council, Durban, South Africa

²University of KwaZulu-Natal, School of Mathematics, Statistics and Computer Science, Pietermaritzburg, South Africa

³Department of Biostatistics, Erasmus University Medical Center, Rotterdam, The Netherlands

Background

HIV remains a major public health challenge in South Africa, particularly among young women and adolescents. Research indicates that sexual health behaviors—including contraceptive use, sexually transmitted infections (STIs), and condom use—do not operate independently but interact in complex ways that influence HIV risk. For example, inconsistent condom use, high STI prevalence, and certain contraceptive methods may increase susceptibility to HIV acquisition. However, traditional statistical approaches often fail to capture these interdependencies over time. Multivariate joint modeling provides an advanced analytical framework for examining these associations, offering a more comprehensive understanding of how sexual behaviors contribute to HIV transmission risk.

Methods

Inspired by the CAP004 clinical trial conducted by CAPRISA which enrolled sexually active, HIV-uninfected women aged 18 to 40 years in South Africa, this study employs multivariate joint modeling to analyze longitudinal data on sexual behaviors (contraceptive use, STIs, and condom use) alongside time-to-event outcomes (HIV infection and pregnancy, used as

a proxy for HIV risk). By simultaneously accounting for correlated processes, this approach enables more accurate estimation of how these behaviors evolve and influence HIV acquisition over time. Data will be drawn from a cohort of individuals at high risk for HIV, with repeated measures of sexual health indicators and HIV outcomes. This modeling framework allows for the identification of trends and causal pathways that conventional regression techniques may overlook.

Results

We anticipate that inconsistent condom use, STI presence, and specific contraceptive methods—such as depot medroxyprogesterone acetate—will be associated with an increased risk of HIV acquisition. Multivariate joint modeling is expected to provide stronger evidence of how these factors interact over time, uncovering key risk patterns. The results will offer deeper insights into the longitudinal dynamics of sexual health behaviors and their contribution to HIV transmission, demonstrating the advantages of this modeling approach over traditional methods.

Conclusion

While the application of joint models in South African HIV research remains limited, studies in other contexts have shown their potential to uncover critical risk factor interactions. This study will contribute novel insights into the complex interplay of sexual health behaviors and HIV risk in South Africa, informing the development of more effective, evidence-based HIV prevention strategies tailored to high-risk populations.

3: Revealing Platelet Aggregation Dynamics: A Functional Data Analysis Approach Using Penalized Splines and Linear Mixed Models.

Souvik Kumar Bandyopadhyay

Cytel, India

The study of platelet aggregation kinetics often requires methods that can capture the full complexity of dynamic processes. Traditional approaches frequently rely on summary statistics, which can obscure critical information embedded within the temporal evolution of the data. This work presents a powerful alternative: a Functional Data Analysis (FDA) approach that models aggregation curves using penalized splines within a Linear Mixed Model (LMM) framework.

Our method leverages the flexibility of Truncated Polynomial Splines (TPS) to represent complex curve shapes. TPS combines polynomial terms and truncated power functions,

enabling smooth curve estimation while penalizing roughness through the LMM structure. By treating polynomial coefficients as fixed effects and truncated terms as random effects, we can utilize Best Linear Unbiased Prediction (BLUP) for robust estimation and analytical derivation of derivatives.

A key innovation of this approach lies in the estimation and interpretation of derivatives, such as velocity and acceleration, which provide critical insights into the underlying kinetics of the system. Unlike traditional methods that focus on static endpoints, these derivatives allow us to model the dynamic system using differential equations, revealing phenomena such as bistability and phase transitions.

We demonstrate the utility of this method through an application to platelet aggregation kinetics, where we analyze the effects of ADP and gold nanoparticles on aggregation profiles. This application showcases how the method can capture the nuances of complex biological interactions, revealing the interplay between ADP concentration, purinergic receptor activation, and nanoparticle-induced effects. Specifically, the method allowed for discerning the underlying kinetics, as well as to study the effects of ADP dosage and perturbation with gold nanoparticles.

This FDA approach, implemented using R with the nlme package, offers significant advantages over traditional methods. By capturing full temporal profiles and identifying metastable states, it provides a more comprehensive understanding of platelet aggregation dynamics. However, it's crucial to address the computational challenges associated with high-dimensional random effects and the sensitivity of derivatives to noise.

This method's generality makes it applicable to other temporal biomedical datasets requiring kinetic analysis. By providing a flexible and statistically rigorous framework for modeling dynamic processes, this work contributes to a deeper understanding of complex biological systems.

4: Source Data Mapping Approach to CDMV5.4: Innovations in Longitudinal Data Integration for Machine Learning Application

Bylkah Mugotitsa^{1,2}, Michael Ochola¹, Pauline Andeso¹, David Amadi³, Reinpeter Momanyi¹, Evans Omondi¹, Jim Todd⁴, Agnes Kiragga¹

¹African Population and Health Research Center, Kenya

²Strathmore Business School, Strathmore University, Nairobi, Kenya.

³Department of Population Health, London School of Hygiene and Tropical Medicine, Lon-

don, United Kingdom

⁴Department of Epidemiology, Catholic University for Health and Allied Sciences Mwanza, Tanzania.

Longitudinal studies offer critical insights into health conditions but face challenges in integrating survey and psychometric data into standardized frameworks like the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). With the cancellation of the survey_conduct table in CDM 6.0, this study proposes a novel methodology to map standardized instruments such as PHQ-9 and GAD-7 into the OMOP CDM, addressing critical gaps in the model.

The methodology leverages a structured process for data integration, starting with the creation of measurements using vocabularies like LOINC and SNOMED. Instrument scores and individual panel items are mapped as observations, with relationships like "Has Answer" ensuring consistency. Tying measurements to visit occurrences through observation_event_id maintains data integrity across longitudinal encounters. Conditions such as "anxiety disorder" are identified and mapped using SNOMED concepts, linked back to corresponding measurements. Insights from OHDSI forums informed iterative refinements, ensuring compatibility with existing vocabularies.

Results show that the methodology effectively integrates psychometric instruments into the OMOP CDM, enabling advanced analysis of mental health outcomes. By addressing the absence of survey_conduct capabilities, the framework facilitates machine learning applications, such as predictive modeling and clustering of longitudinal data, while preserving the semantic integrity of health data.

This work provides a scalable solution for mapping survey data into the OMOP CDM, bridging a key gap in longitudinal data management and advancing global health research. The framework enhances data standardization and usability, contributing to evidence-based policy-making and interventions. Future efforts will extend this methodology to emerging instruments and vocabularies, supporting the evolution of data science frameworks in health research.

Biomarker Studies & Mixed Topics

Wednesday, 2025-08-27 09:00 - 10:30, ETH E21

Chair: Krista Fischer

1: Quantifying the Clinical Usefulness of Novel Biomarkers and Tests: Beyond Traditional Statistics

Frank Doornkamp¹, Jelle J Goeman¹, Ewout W Steyerberg^{1,2}

¹LUMC, Netherlands, The

²UMCU, Netherlands, The

Background New prognostic tests and biomarkers are often described with sensitivity and specificity, but evaluating their clinical usefulness requires moving beyond traditional performance metrics. We aim to evaluate how new biomarkers improve health outcomes through improved treatment allocation, compared to current standard care.

Methods We developed a decision analytic model comparing treatment decision making based on a reference prediction model versus the reference model combined with a new biomarker. Risk distributions for a clinical outcome were simulated based on a reference model alone or with the new biomarker. Individual treatment benefit was estimated assuming a constant relative effect. Treatment was recommended if the absolute risk reduction exceeded a defined treatment threshold. Traditional statistical performance measures included sensitivity, specificity, and the area under de ROC curve (AUC). The clinical usefulness of the biomarker was quantified with Net Benefit: a weighted sum between the reduction in events and the number of treatments given, per 10,000 patients. Uncertainty of the improvement in Net Benefit was assessed by subsampling from the simulated data set. As an illustrative example, we assessed the clinical usefulness of the genomic MammaPrint test (assumed sensitivity 64%, specificity 67%) in addition to the standard clinical risk assessment (PREDICT model, <https://breast.predict.cam/tool>) for early breast cancer patients. Systematic sensitivity analyses assessed the drivers of clinical usefulness.

Results For early breast cancer, adding the MammaPrint to the PREDICT model increased the AUC from 0.69 to 0.72. Net Benefit increased from 13 net distant metastases prevented to 15 per 10,000. Substantial uncertainty was noted by drawing samples equal to the MINDACT trial, the study that provided key evidence on the incremental value of MammaPrint test (n=6693), suggesting further evidence is needed to claim clinical usefulness. Sensitivity analysis showed that the clinical usefulness of a new test was larger if the event rate was

higher, treatment effect larger, its own quality better, or quality of the reference model lower. These patterns were not consistent by increases in AUC. We present test and context characteristics in a ShinyApp to facilitate early assessments of the potential clinical usefulness of novel tests and biomarkers.

Conclusions Decision analytic modeling provides insights into how the sensitivity and specificity of a new biomarker translate to clinical usefulness within its clinical context. We found that clinical usefulness depends not only on its prognostic strength but also on key contextual factors, which are not captured by traditional statistical performance measures.

2: The Underlap Coefficient as Measure of a Biomarker's Discriminatory Ability in a Multi-Class Disease Setting

Zhaoxi Zhang, Vanda Inácio, Miguel de Carvalho

University of Edinburgh (United Kingdom)

Background The first step in evaluating a potential diagnostic biomarker is to examine the variation in its values across different disease groups. The significance of employing appropriate metrics in the evaluation phase cannot be overstated. The most commonly used metrics for this purpose are predominantly Receiver Operating Characteristic (ROC) based. However, these measures rely on a stochastic ordering assumption for the distributions of the biomarker's outcomes across groups. This assumption can be restrictive, particularly when covariates are involved, and its violation may lead to incorrect conclusions about a biomarker's ability to distinguish between disease classes. Even when a stochastic ordering exists, the order may vary across different biomarkers in discovery studies, complicating automated ranking.

Methods To address these challenges and complement existing measures, we propose the underlap coefficient (UNL), a novel summary index of a biomarker's ability to distinguish between multiple disease groups, and study its properties particularly in the three-class case. We establish a direct analytical link between the UNL and the three-class Youden index (YI), as well as between the UNL and the Weitzman's two-class overlap coefficient (OVL). These relationships can be easily generalized to settings with more than three disease classes. We also numerically explore the relationship between the UNL and the volume under the ROC surface (VUS) in a proper trinormal framework. Additionally, we introduce Bayesian nonparametric estimators for both the unconditional underlap coefficient and its covariate-specific counterpart. Furthermore, we illustrate the proposed approach through an application to an Alzheimer's disease (AD) dataset aimed to assess how four potential AD biomarkers

distinguish between individuals with normal cognition, mild impairment, and dementia, and how and if age and gender impact this discriminatory ability.

Results A simulation study reveals a good performance of the proposed estimators across a range of conceivable scenarios. The results from the application study indicate a moderate age and gender effect on the biomarkers' discriminatory ability. Also, by comparing UNL with the three-class YI in the application study, we find that YI could be underestimating some biomarkers' discriminatory capability at certain covariate values.

Conclusion We discuss the underlap coefficient as a measure of diagnostic accuracy in a multi-class disease framework. It offers advantages over ROC-based summary measures for evaluating the diagnostic potential of a biomarker during its discovery phase, as it does not require an assumed order of classes and is better suited for multi-modal density settings.

3: Time-Dependent Accuracy for Continuous Biomarkers using Copula Modelling

Adina Najwa Kamarudin¹, Ahmad Faiz Mohd Azhar¹, Nurain Ibrahim²

¹Universiti Teknologi Malaysia, Malaysia

²Universiti Teknologi MARA, Malaysia

In biomedical research, a key interest lies in the building of classification models to analyse patient survival based on biomarkers, with patients being discriminated into different cases. If a biomarker is dependent on time, the accuracy of these models can be assessed through the time-dependent receiver operating characteristic (ROC) curve. The sensitivity and specificity of the classification model are measured by this curve to detect which patients have long or short survival times. The accuracy trend is produced by computing the area under this curve (AUC) at each time point. Several methodologies have been proposed for the estimation of the accuracy trend, revolving around nonparametric or semiparametric estimators. The use of these methods may limit researchers from suggesting additional reasons for why accuracy measurements are high or low at certain time points. In this paper, a demonstration of how the accuracy trend of a classification model can be estimated parametrically from a time-dependent ROC curve is provided. Based on a simulation study on copula functions, it is shown that the accuracy value and trend may be influenced by the dependence measurements or the selection of copulas. This can improve the understanding of the accuracy trend of a classification model. In a real application, the statistical information of a single biomarker/score derived from the primary biliary cholangitis (PBC) dataset was linked to its time-to-event using the Gaussian copula. It was observed that the biomarker/score derived from five covariates gives the highest accuracy performance, as the strongest negative

dependence structure was found compared to other markers.

Keywords time-dependent, AUC, copula, sensitivity, specificity.

4: Sample Size Determination for Hypothesis Testing of the Intraclass Correlation Coefficient for Agreement in Two-Way ANOVA Models

Dipro Mondal¹, Alberto Cassese², Math JJM Candel¹, Sophie Vanbelle¹

¹Department of Methodology and Statistics, Care and Public Health Research Institute (CAPHRI), Maastricht University, The Netherlands

²Department of Statistics, Computer Science, Applications "Giuseppe Parenti", The University of Florence, Italy

Introduction Reliability assessment is essential in medical domains to ensure accurate patient diagnosis. When multiple raters evaluate the same patients using quantitative measurements, a two-way ANOVA model may be appropriate, with the intraclass correlation coefficient for agreement (ICCa) serving as the reliability metric.

Designing such reliability studies requires determining the number of patients and raters. While sample size procedures exist based on the expected width of confidence intervals for ICCa, procedures based on hypothesis testing remain underdeveloped. These procedures utilise the lower limit of the confidence interval for ICCa [1, 2] and determine sample sizes ensuring adequate power for testing whether ICCa exceeds a predefined threshold. We propose sample size procedures for hypothesis testing building on available confidence interval methods for ICCa.

Methods We identify seven classes of confidence interval methods for ICCa and compare their empirical type-I error rates. Focusing on the best performing methods, simulation-based sample size determination procedures are proposed. These procedures are evaluated by assessing the empirical power of the hypothesis test at the calculated sample size. Accessibility of these procedures is facilitated by implementing an interactive R/Shiny app.

Results Comparison of the type-I error rates of the confidence interval methods indicates that the rater-to-error variance ratio influences which method emerges as the best-performing in maintaining the type-I error rate close to the nominal value. Evaluation of our proposed sample size procedures shows that they provide adequate power across most parameter configurations.

Conclusion The rater-to-error variance ratio should guide practitioners in selecting an appropriate confidence interval method for ICC_a. Our proposed sample size procedures, along with the R/Shiny app implementation, provide a practical framework for designing reliability studies.

1. Mondal, D., et al., *Review of sample size determination methods for the intraclass correlation coefficient in the one-way analysis of variance model*. Statistical Methods in Medical Research, 2024. **33**(3): p. 532-553.
2. Zou, G.Y., *Sample size formulas for estimating intraclass correlation coefficients with precision and assurance*. Stat Med, 2012. **31**(29): p. 3972-81.

5: Accounting for Misclassification of Binary Outcomes in External Control Arm Studies for Unanchored Indirect Comparisons: Simulations and Applied Example

Mikail Nourredine^{1,2}, Antoine Gavoille^{1,2}, Côme Lepage^{3,4}, Behrouz Kassai-Koupai^{2,5}, Michel Cucherat⁶, Fabien Subtil^{1,2}

¹Service de Biostatistique-Bioinformatique, Hospices Civils de Lyon, F-69003 Lyon, France

²Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Évolutive UMR 5558, F-69100 Villeurbanne, France

³Fédération Francophone de Cancérologie Digestive, EPICAD INSERM UMR CTM 1231, University of Burgundy and Franche Comté, Dijon, France

⁴Department of Digestive Oncology, University Hospital Dijon, University of Burgundy and Franche Comté, Dijon, France

⁵Service hospitalo-universitaire de pharmaco-toxicologie de Lyon, Centre d'Investigation Clinique CIC 1407, Inserm-Hospices Civils de Lyon, Lyon France

⁶Service de pharmacologie et de toxicologie (metaEvidence.org), Hospices Civils de Lyon, Lyon, France

Introduction Statistical methods for single-arm trials with an External Control Arm (ECA) usually assume no difference in outcome measurement between arms. However, ECA data may measure only proxy outcomes, leading to potential misclassification and biased estimates. This study aimed to quantify bias from ignoring binary outcome misclassification and propose a likelihood-based correction method.

Methods The proposed model relies on a validation study in which both proxy and reference outcomes are measured, to overcome the misclassification problem, and a joint likelihood estimation for the validation, ECA, and the prospective single-arm data. In addition to stan-

dard assumptions in indirect treatment comparisons, it requires a correctly specified outcome measurement error model that accounts for all variables contributing to non-differential measurement error. We performed simulations varying sample size, specificity, and sensitivity, evaluating relative bias, empirical standard error (SE), root mean square error (RMSE), and 95% confidence interval (CI) coverage. In an applied example, we compared sorafenib with a proxy outcome (PRODIGE11-trial) versus placebo with the reference outcome (SHARP-trial), using the SHARP trial's gold standard treatment effect estimate.

Results Simulations showed that ignoring misclassification in binary outcomes leads to substantial bias in the estimation of indirect treatment effects. Even with a specificity and sensitivity at 0.9, the uncorrected method had a relative bias of 67%. The proposed model reduced bias in all simulation sets, with a relative bias below 5% and a 95%CI coverage between 95% and 96.5%. Across varying levels of specificity and sensitivity, the proposed method achieved approximately half the RMSE of the uncorrected method. Additionally, increasing the ECA sample size had a greater impact on reducing the proposed method's RMSE than enlarging the validation study sample size. The gold standard effect of sorafenib compared with placebo was OR=0.52 (SHARP). Ignoring outcome misclassification resulted in an overestimation of the indirect treatment effect (OR=0.36), using the proposed model the estimation was OR=0.55. However, with only 161 patients in the sorafenib arm of the PRODIGE-11 trial, the 95%CI estimated by the proposed model was wide. This is conservative, as it transfers the uncertainty in measurement to the uncertainty in the decision. These findings align with simulation results, where the empirical SE of the proposed model was twice that of the reference outcome regression for a sample size of 200 patients.

Conclusions The findings underscore the importance of addressing outcome misclassification in indirect comparisons. The proposed correction method may improve reliability in unanchored indirect treatment comparisons.

Randomization and Analysis of Clinical Trials

Wednesday, 2025-08-27 14:00 - 15:30, Biozentrum U1.131

Chair: Stephen Senn

1: Distributive Randomization: a Pragmatic Design to Evaluate Multiple Simultaneous Interventions in a Clinical Trial

Skerdi Haviari^{1,2}, France Mentré^{1,2}

¹Université Paris Cité, Inserm, IAME, 75018 Paris, France

²Département Epidémiologie Biostatistiques Et Recherche Clinique, AP-HP, Hôpital Bichat, 75018 Paris, France

Background In some medical indications, numerous interventions have a weak presumption of efficacy, but a good track record or presumption of safety. This makes it feasible to evaluate them simultaneously. This study evaluates a new design that randomly allocates a pre-specified number of interventions to each participant, and statistically tests main intervention effects. We compare it to factorial trials, parallel-arm trials and multiple head-to-head trials, and derive some good practices for its design and analysis. We extend the approach by varying the number of interventions from one patient to the next for the simultaneous evaluation of an intervention program, comprising several components, and each component individually, at the same time, enabling proper contrasts.

Methods We simulated various scenarios involving 4 to 20 candidate interventions/components among which 2 to 8 could be simultaneously allocated. A binary outcome was assumed. One or two interventions were assumed effective, with various interactions (positive, negative, none). Efficient combinatorics algorithms were created. Sample sizes and power were obtained by simulations in which the statistical test was either difference of proportions or multivariate logistic regression Wald test with or without interaction terms for adjustment.

Results Distributive trials reduce sample sizes 2- to 7-fold compared to parallel arm trials. An unexpectedly effective intervention causes small decreases in power (< 10%) if its effect is additive, but large decreases (possibly down to 0) if not, as for factorial designs. These large decreases are prevented by using interaction terms to adjust the analysis, but these additional estimands have a sample size cost and are better pre-specified. The issue can also be managed by adding a true control arm without any intervention, or by exploiting the variance-covariance matrix, which is all the more useful for the multi-component intervention use case.

Conclusion Distributive randomization is a viable design for mass parallel evaluation of interventions in constrained trial populations. It should be introduced first in clinical settings where many undercharacterized interventions are potentially available, such as disease prevention strategies, digital behavioral interventions, dietary supplements for chronic conditions, or emerging diseases. Pre-trial simulations are recommended, using publicly available code.

2: Forced Randomisation – a Powerful, Sometimes Controversial, Tool for Multi-Centre RCTs

Johannes Krisam¹, Kerstine Carter², Olga Kuznetsova³, Volodymyr Anisimov⁴, Colin Scherer⁵, Yevgen Ryeznik⁶, Oleksandr Sverdlov⁷

¹Boehringer Ingelheim Pharma GmbH & Co.KG (Ingelheim, Germany)

²Boehringer Ingelheim Pharmaceuticals Inc. (Ridgefield, CT, USA)

³Merck & Co. Inc. (Rahway, NJ, USA)

⁴Amgen Ltd. (London, United Kingdom)

⁵Rensselaer Polytechnic Institute (Troy, NY, USA)

⁶Uppsala University (Uppsala, Sweden)

⁷Novartis Pharmaceuticals Corporation (East Hanover, NJ, USA)

During the enrolment period of a randomised controlled trial, drug supply at a site may run low, such that a site does not have medication kits from all types available. In case an eligible patient is to be randomised to a treatment for which no kits are currently available, two options are possible: Either send that patient home, which might be deemed as ethically questionable; Or allocate the patient to a treatment arm with available kits at the site, using a built-in feature of the interactive response technology (IRT) system, called forced randomisation (FR). In the pharmaceutical industry, there is a general consensus that using FR might be acceptable, given that there are “not too many” instances of FR. Furthermore, FR could be considered at odds with the ICH E9 guidance [1], which states that “[t]he next subject to be randomised into a trial should always receive the treatment corresponding to the next free number in the appropriate randomisation schedule (in the respective stratum, if randomisation is stratified)”. Unfortunately, a clear guidance on under what instances FR is acceptable is currently lacking.

This talk will present recent work covering the potential benefits that can be garnered from the use of forced randomisation under various forcing, and supply strategies [2]. The impact on important characteristics of the clinical trial, such as the balance in sample size between treatment arms, the number of patients sent home, the duration of the trial as well as the drug

overage will be discussed. In addition, potential ways on how to address forced randomisation in the statistical analysis will be assessed, and the impact of forced randomisation on the type I error rate of a trial will be investigated under several scenarios.

1. ICH Harmonised Tripartite Guideline E9. Statistical Principles for Clinical Trials (1998).
https://database.ich.org/sites/default/files/E9_Guideline.pdf
2. Carter K, Kuznetsova O, Anisimov V, Krisam J, Scherer C, Ryeznik Y, Sverdlov O (2024). Forced randomization: the what, why, and how. BMC Med Res Methodol 24(1):234. doi: 10.1186/s12874-024-02340-0.

3: Type I Error Rate Control in Adaptive Platform Trials when Including Non-Concurrent Controls in the Presence of Time Trend

Jinyu Zhu¹, Peter Kimani¹, Nigel Stallard¹, Andy Metcalfe¹, Jeremy Chataway², Keith Abrams¹

¹University of Warwick, United Kingdom

²University College London, United Kingdom

Background Platform randomized controlled trials (RCTs) have gained increased popularity during and after pandemic due to their efficiency and resource-saving capabilities. These trials allow the addition or removal of experimental treatments at any stage. Usually, only concurrent controls are used when an added experimental treatment arm is compared to the control arm. However, platform trials offer the opportunity to include non-concurrent controls, which are controls recruited before the added experimental arm entered the trial. Using non-concurrent controls increases power but requires considering time trends because ignoring time trends introduces bias in making inference on treatment effects.

Methods Lee and Wason (2020) proposed fitting a linear model with terms for time trend, which was later extended by Bofill Roig et al. (2022). This, however, does not account for interim analyses associated with adaptive platform RCTs. We build on this model to show how to compute the interim and final analyses boundaries for testing the effect of an added treatment that control the type I error rate. We first demonstrate that using a linear model with additional parameters for added treatments as the RCT progresses, the test statistics for testing a treatment effect at different interim analyses, have the joint canonical distribution assumed in group sequential methods (Jennison and Turnbull, 1997). Then we identify boundaries that control the type I error rate for any values of the true effects of experimental arms introduced earlier in the trial.

Results Our findings show that borrowing information from non-concurrent controls can improve power, though it depends on possible decisions at interim analyses (futility stopping only, efficacy stopping only or both) and effects of the experimental treatments that entered the trial earlier. Also, the boundaries can be optimized in terms of power.

Conclusion This study establishes a method to control type I error while optimising power in platform RCTs with non-concurrent controls, accounting for both time trend and interim analyses.

Reference Bofill Roig et al., 2022. On model-based time trend adjustments in platform trials with non-concurrent controls. *BMC medical research methodology*, 22(1), p.228.

Jennison and Turnbull, 1997. Group-sequential analysis incorporating covariate information. *Journal of the American Statistical Association*, 92(440), pp.1330-1341.

Lee and Wason, 2020. Including non-concurrent control patients in the analysis of platform trials: is it worth it?. *BMC medical research methodology*, 20, pp.1-12.

4: Data-Driven Controlled Subgroup Selection in Clinical Trials

Manuel M. Müller¹, Konstantinos Sechidis², Björn Bornkamp², Frank Bretz², Fang Wan³, Wei Liu⁴, Henry W. J. Reeve⁵, Timothy I. Cannings⁶, Richard J. Samworth¹

¹Statistical Laboratory, University of Cambridge, Cambridge, United Kingdom

²Advanced Methodology and Data Science, Novartis Pharma AG, Basel, Switzerland

³Department of Mathematics and Statistics, Lancaster University, Lancaster, United Kingdom

⁴School of Mathematical Sciences, University of Southampton, Southampton, United Kingdom

⁵School of Mathematics, University of Bristol, Bristol, United Kingdom

⁶School of Mathematics, University of Edinburgh, Edinburgh, United Kingdom

Background and Introduction

Subgroup selection in clinical trials is essential for identifying patient groups that may benefit differently from a treatment, thereby enhancing personalized medicine. Additionally, it can identify patient groups that encounter adverse events after treatment. However, these post-selection inference problems pose challenges, such as increased Type I error rates and potential biases from data-driven subgroup identification. In this paper, we present and extend two recently developed tools for subgroup selection in regression problems: one based

on generalized linear modelling (GLM) (<https://doi.org/10.1002/sim.9996>) and another on nonparametric monotonicity constraints (<https://doi.org/10.1093/jrsssb/qkae083>). These methods alleviate the above reliability concerns and we demonstrate how these methods can be extended and applied to address questions regarding treatment effect heterogeneity.

Methods

To evaluate these methods' effectiveness and reliability in clinical settings, we conduct a thorough simulation study in which the data distributions mimic those of real data sets; one of which is observational data and one data from a randomized controlled trial. Finally, we apply the methods to the data of the original two clinical trials to address two distinct questions: identifying patient groups that manifest adverse events and identifying patient groups that experience enhanced treatment effects, while controlling for Type I error in both cases.

Results.

We assess how well the methods retain Type I error rate control under violation of their respective assumptions. We find that while the GLM-based method is less robust against such violations compared to the monotonicity-based approach, when its underlying modelling assumptions hold it has higher power. However, a more fine-grained analysis suggests that while the very strict notion of Type I error is violated more easily in the parametric setting, the story is less clear for nuanced measures of reliability. Furthermore, we illustrate the suitability of the examined methods when applied on top of the meta-learning framework popular in evaluating conditional average treatment effects.

Conclusions.

We conclude that recent methods for controlled subgroup selection exhibit a trade-off in their reliability and power which parallels that between parametric or nonparametric methods elsewhere in statistics. Our study investigates the extend of these effects, which should serve as useful guidance for using controlled subgroup selection approaches in other applications. Of particular interest is the difference between the different measures of reliability we consider, the appropriate choice of which will strongly depend on the real-world application at hand.

5: Debunking the Myth: Random Block Sizes Do not Decrease Selection Biases in Open-Label Clinical Trials

Wenle Zhao

Medical University of South Carolina, United States of America

Background The permuted block design for subject randomization in clinical trials is highly susceptible to selection biases due to its predictability patterns, particularly when investigators are aware of the block size. This predictability can influence their enrollment decisions, introducing selection biases, especially in open-label trials. The random block design aims to mitigate this issue by incorporating randomly varying block sizes, expecting that without knowing the block size, investigators will be less likely to make treatment predictions. For this reason, it has been recommended by the ICH E9 Statistical Principles for Clinical Trials. However, close examinations revealed that 100% certainty is not a necessary condition for making treatment predictions; any prediction with correct guess probability above pure random can result in selection biases.

Methods Our recent research provides an analytical framework for the assessment of allocation predictability for infinite and finite allocation sequences. The convergence guessing strategy, proposed by Blackwell and Hodges in 1957, offers a more realistic measure of selection bias compared to prediction with certainty. Using the correct guess probability as a metric, we evaluated the selection bias of random block design and compare it with the permuted block design and alternative randomization designs, all under the same restriction of maximum tolerated imbalance.

Results Quantitative assessments indicate that, among all restricted randomization designs with the same restriction of maximum tolerated imbalance, the random block design exhibits the highest risk of selection bias. For instance, in a two-arm equal allocation trial with the maximum tolerated imbalance of 3 (block size of 6), the average selection bias risk is 68.33% for the permuted block design and 70.28% for the random block design, while the big stick design and the block urn design have 58.23% and 62.35% respectively.

Conclusion Replacing permuted block design with random block design increases the risk of selection biases and should not be recommended to use in open label trials. Instead, the big stick design and the block urn design offer superior protection against selection bias and therefore are recommended to be used in open label trials.

References

1. Zhao W, Carter K, Sverdlov O, et al. Steady-state statistical properties and implementation of randomization designs with maximum tolerated imbalance restriction for two-arm equal allocation clinical trials. *Stat Med*. 2024;43(6):1194-1212.
2. Zhao W, Livingston S. Allocation Predictability of Individual Assignments in Restricted Randomization Designs for Two-Arm Equal Allocation Trials. *Stat Med*. 2025;44(3-4):e10343.

Survival Analysis 2

Wednesday, 2025-08-27 14:00 - 15:30, Biozentrum U1.141

Chair: Martina Mittlböck

1: A Framework for Estimating, Investigating and using the Correlation Between Multiple Time-to-Event Endpoints in a Group Sequential Trial

Anne Lyngholm Soerensen^{1,2}, Paul Blanche¹, Henrik Ravn², Christian Pipper²

¹University of Copenhagen (Denmark)

²Novo Nordisk (Denmark)

Introduction A correlation estimate is used to calculate the expected power of rejecting hypotheses related to multiple endpoints during and at the end of a group sequential trial (GST). Thus, determining an appropriate correlation between endpoints is key during the planning of the trial. This entails using data from earlier comparable trials. However, when endpoints are time-to-event several challenges present themselves. In particular, the censoring scheme of a given trial will influence the correlation between commonly used log-rank test-statistics and in general the correlation is not expected to adhere to any simple canonical form. We will present a method to estimate the time-dependent correlation of test statistics for time-to-event endpoints. The method will further allow for investigation of what drives the correlation and how its input can be modified to aid in the design of future trials or GSTs with time-to-event endpoints.

As the correlation can be used for more efficient testing, the method can provide an estimate of the correlation to provide more powerful confirmatory testing strategies.

Methods Using the identical and independent distributed (iid) decomposition of the log-rank test statistics for the endpoints, we can calculate the time-dependent correlation in previously trials. The decomposition allows us to understand what drives the correlation. We can visibly in the expression of the decomposition isolate several operational characteristics such as the time of censoring, the timing of inclusion, and more. By altering the characteristics, we can estimate how they affect the correlation between the endpoints during the GSTs and identify main drivers of the correlation. This creates a plug-and-play tool for using information from earlier trials to be adapted with the design decisions and expectations from new trials. It is then possible to plan future trials via the method and simulation without enforcing assumptions about the correlation structure.

The implementation of the method and how it can be used for estimating, investigating and planning future trials will be shown using data from a previous cardiovascular outcomes trial.

Results The iid decomposition of the log-rank test statistics in a time-to-event trial provides a powerful tool for estimating, investigating and using the correlation between multiple time-to-event endpoints in a GST.

2: Likelihood Adaptively Incorporated External Aggregate Information with Uncertainty for Survival Data

Jing Ning

The University of Texas M.D. Anderson Cancer Center, United States of America

Introduction Population-based cancer registry databases are invaluable for bridging the gap created by the limited statistical power of primary cohort studies with small to moderate sample sizes. While these databases often lack detailed tumor biomarker data or report it inconsistently, they provide comprehensive and publicly accessible aggregate survival statistics. Integrating such data with primary cohorts holds promise for enhancing treatment evaluation and survival prediction across tumor subtypes. However, in rare cancers, even registry sample sizes may be modest, and the variability associated with aggregated statistics can be substantial relative to the primary cohort's sample variation. Neglecting this variability risks misleading conclusions.

Methods We propose a likelihood-based method that adaptively incorporates external aggregate information while accounting for its variability. To ensure computational efficiency and stability, we introduce a nuisance parameter to circumvent the infinite-dimensional baseline hazard function in aggregate data. We derive the asymptotic properties of the estimators and assess their finite-sample performance through simulations.

Results Simulation studies demonstrate that the proposed method performs robustly across varying levels of external information variability, outperforming existing approaches. We applied the method to integrate inflammatory breast cancer (IBC) patient data from the University of Texas MD Anderson Cancer Center with aggregate survival data from the National Cancer Data Base. This enabled an assessment of the Ki-67 biomarker (negative vs. positive) in predicting the survival benefits of trimodality treatment. Results indicated poorer survival outcomes for Ki-67 positive patients, characterized by higher cancer cell proliferation, compared to their negative counterparts with lower proliferation. Trimodality

treatment significantly benefited Ki-67 positive patients, while Ki-67 negative patients derived limited survival benefit. These findings highlight Ki-67's potential as a predictive biomarker for tailoring therapy in IBC.

Conclusion In real-world data integration, accounting for the variability in aggregate information is critical to avoid bias and enhance statistical efficiency. Our method appropriately incorporates external data variability, safeguarding against the integration of unsuitable external information due to population heterogeneity. By ensuring data-driven borrowing of information, the approach enhances inference accuracy and supports informed decision-making in precision oncology.

3: A New Statistical Test to Compare Probability of Being in Response (PBR) with Application to a Study in Oncology

Norbert Hollaender¹, Ekkehard Glimm^{1,2}

¹Novartis Pharma AG, Basel, Switzerland

²Otto-von-Guericke University, Institute of Biometry and Medical Informatics, Magdeburg, Germany

INTRODUCTION

The probability-of-being-in-response (PBR) function provides easily interpretable curves for time from treatment start to first response and time from first response to subsequent failure. Comparison between treatment arms is based on visual inspection of the PBR curves, inference is rarely applied in practice. Here, we describe a statistical test for a comparison of PBR curves.

METHODS

The PBR function can be derived from a multistate model. At study start, all patients are in an initial state 0 (not in response). Patients responding to treatment enter state 1 (in response) at time of first documented response. Later, they might enter the absorbing state, state 2. For comparison of two PBR curves we consider three test statistics which are extensions of the logrank test for right censored survival curves. The derivation of the test statistics' distribution is based on conditional probabilities of entering the response state given the risk sets in the treatment arms at each event time. In addition, the risks sets for leaving the response state at the event times are also considered. We describe the statistical methodology, investigate type-I errors and power in a simulation study and illustrate the

application using data from the clinical phase 3 study REACH3.

RESULTS

The suggested tests keep the type I error under a Markov property. Simulations show that i) high power is achieved for event rate ratios (of the event types 'entering state 1' and 'leaving state 1') between the treatments that are constant in time and the test treatment is the better one, ii) the type I error is preserved if the test treatment is consistently no better than the control treatment over the entire time axis and iii) if event rate ratios are above 1 for some time periods and below 1 for others, there are no statistical guarantees of the tests' properties. For REACH3, the tests confirm statistical significance of the observed differences between PBR curves.

CONCLUSION

The proposed tests are straightforward extensions of the logrank test. Simulations and the application to clinical trial data show that these tests are useful additions to the visual comparison of PBR curves.

4: One-Sample Survival Tests for Non-Proportional Hazards in Oncology Clinical Trials

Chloé Szurewsky¹, Guosheng Yin², Gwénaël Le Teuff¹

¹CESP, INSERM U1018, University of Paris-Saclay, France

²Department of Statistics and Actuarial Science, University of Hong-Kong Pokfulam Road, Hong Kong

In oncology, well-powered time-to-event randomised clinical trials are challenging for rare diseases (e.g., pediatric cancers or personalised medicine) due to limited patient numbers. One- or two-stage designs for single-arm trials (SATs) with time-to-event outcomes have emerged in recent years as compelling alternatives to overcome this issue. These designs rely on the one-sample log-rank test (OSLRT) and its modified version (mOSLRT) to compare the survival curve of an experimental arm to that of an external (or historical) control group under the proportional hazards (PH) assumption that may be violated particularly when evaluating immunotherapies. We develop score tests and investigate alternatives for situations where PH does not hold. We extend Finkelstein's score test (OSLRT) developed under PH by using a piecewise-exponential (PE) model with change-points (CPs) for early, middle and delayed treatment effect and an accelerated hazards model for crossing hazards. The restricted

mean survival time (RMST-) based test is adapted to the case of SATs. We construct a maximum combination (max-Combo) test combining the mOSLRT, early and delayed score tests. The performances (type I error and power) of the developed tests are evaluated through a simulation study. The survival times are generated with an exponential distribution assuming no sampling variability for the reference group and a PE for the experimental group. The simulation parameters are the sample size of the experimental group (from 20 to 200 patients), the exponential censoring rate (from 0 to 35%) and the relative treatment effect (hazard ratio from 0.5 to 1). A SAT in paediatric patients with neuroblastoma evaluating an inhibitor is used for illustration. The simulation study shows that the score tests are more conservative than the mOSLRT and as conservative as the OSLRT. The score test has the highest power when the data generation matches with the model even when CPs are misspecified. The RMST-based test is less powerful than the mOSLRT except for an early effect with censoring rate less than 15%. The max-Combo test is conservative and more powerful than the mOSLRT with sufficient sample sizes ($n > 50$) but less than the appropriate score test under non-PH. To conclude, the score tests are efficient under non-PH when the approximate values of the CPs are known a priori and the max-Combo is an alternative when the time-dependant treatment effect and values of CPs are unknown. Further researches needs to be conducted to study the impact of the external control group sampling variability.

5: Efficiency of Nonparametric Superiority Tests Based on Restricted Mean Survival Time Versus the Logrank Test Under Proportional Hazards

Dominic Magirr

Novartis, Switzerland

For randomized clinical trials with time-to-event endpoints, proportional hazard models are typically used to estimate treatment effects and log-rank tests are commonly used for hypothesis testing. The summary measure of the primary estimand is frequently a hazard ratio. However, there is growing support for replacing this approach with a model-free summary measure and assumption-lean analysis method—a trend already observed for continuous and binary endpoints. One alternative is to base the analysis on the difference in restricted mean survival time (RMST) at a specific timepoint, a single-number summary measure that can be defined without any restrictive assumptions on the outcome model. In a simple setting without covariates, an assumption-lean analysis can be achieved using nonparametric methods such as Kaplan-Meier estimation. The main advantage of moving to a model-free summary measure and assumption-lean analysis is that the validity and interpretation of conclusions do not depend on the proportional hazards (PH) assumption. The potential disadvantage

is that the nonparametric analysis may lose efficiency under PH. There is disagreement in recent literature on this issue, with some studies indicating similar efficiency between the two approaches, while others highlight significant advantages for PH models. We present asymptotic results and a simulation study to clarify the conflicting results from earlier research. We characterize those scenarios where relative efficiency is close to one, and those where it isn't. Several illustrative examples are provided.

Causal Inference: Mixed Topics

Wednesday, 2025-08-27 14:00 - 15:30, Biozentrum U1.101

Chair: Stijn Vansteelandt

1: Establishing when to Use Causal Machine Learning for Conditional Average Treatment Effect Estimation in Randomised Controlled Trials using Simulation

Eleanor Van Vogt¹, Suzie Cro¹, Karla Diaz-Ordaz²

¹Imperial College London, United Kingdom

²University College London, United Kingdom

Background Randomised controlled trials (RCTs) typically focus on estimating the average treatment effect (ATE), which often results in null conclusions. However, where heterogeneous treatment effects (HTEs) are of interest, factors responsible for variation must be pre-specified. There is growing interest in exploring HTEs in the context of personalised treatment regimens and policy decisions, and causal machine learning methods for HTEs are increasing in popularity. They offer flexible tools for exploring HTEs across many covariates without needing pre-specification by learning the conditional average treatment effect (CATE).

Current usage of these methods is restricted to exploring HTEs and generating hypotheses for validation in a future dataset. Existing questions relate to the sample sizes required to obtain valid CATEs and the impact of missing covariate information on resulting CATEs.

Methods We conduct a simulation study to compare several causal machine learning candidates and classical subgroup detection methods across simple and complex HTE scenarios with varying sample sizes and missing data mechanisms. We consider binary and continuous outcomes and the handling of competing events. We explore bias, coverage of HTE estimates, and error rates for global heterogeneity tests.

Informed by minimum sample size requirements from our simulation and results from heterogeneity testing, we additionally simulate scenarios where HTE hypotheses are generated during an interim analysis of an RCT and then validated on the later recruited participants. Large RCTs could potentially use this approach to generate and validate subgroup findings and provide treatment recommendations.

Expected Results and Discussion: By addressing challenges such as minimum sample size

requirements and missing data handling, the presented results from simulations will provide researchers with a framework to decide whether causal machine learning methods are suitable for RCT datasets at their disposal. Further, the proposed interim analysis framework has the potential to enhance RCT utility, enabling real-time hypothesis generation and validation for personalised, evidence-driven treatment.

2: Variable Selection in Causal Survival Analysis

Charlotte Voinot^{1,2}, Julie Josse¹, Bernard Sébastien²

¹Premedical, INRIA, France

²Sanofi R&D, France

Background In classical causal inference, it is well-established that instrumental variables should not be included in adjustment models, whereas precision variables should be included as they lead to a gain in variance, even when employing weighting estimators (IPW). However, no analogous guidelines exist in causal survival analysis, where numerous estimators rely on different nuisance models, yet variable selection strategies remain largely unexamined. Given the various estimators available for restricted mean survival time (RMST), understanding the role of different types of variables—including precision variables, instrumental variables, and censoring-related variables—is crucial for improving estimation efficiency.

Methods We estimate RMST using different causal survival estimators such as the G-formula, IPTW-Buckley-James, IPTW-IPCW Kaplan-Meier, and the doubly/triply robust AIPTW-AIPCW and analyze the impact of variable selection across treatment, outcome, and censoring models. Our study assesses how different types of variables—precision variables (affecting only the outcome), instrumental variables (affecting only treatment), and censoring-related variables—influence estimator variance. In particular, we examine the inclusion of variables that affect both censoring and outcome, which may differ from classical confounders. Variance calculations and simulations are conducted to evaluate the effects of these choices.

Results Our findings confirm that including precision variables in the outcome model improves estimator efficiency. Moreover, consistent with findings in classical causal inference, precision variables also provide benefits when incorporated into the treatment model, even in the context of weighting estimators. Similarly, instrumental variables increase variance in the treatment model, aligning with findings from classical causal inference, reinforcing the necessity of careful variable selection. Regarding censoring-related variables, we find that those affecting both censoring and the outcome improve precision when included in the out-

come model, whereas variables solely related to censoring increase variance and should not be included in the censoring model.

Conclusion This study provides practical recommendations for variable selection in causal survival analysis. Our results highlight that precision variables should be included, while instrumental variables should be avoided. Additionally, we provide new insights specific to survival analysis: censoring-related variables negatively impact variance in the censoring model, but variables influencing both censoring and the outcome enhance precision when included in the outcome model. These findings apply to parametric and semi-parametric models. However, in nonparametric approaches like causal forests, the conclusions may be more nuanced. With finite sample sizes, including additional variables could introduce a bias-variance tradeoff, and the observed benefits in this study would likely hold only asymptotically.

3: An Overlooked Stability Property of the Risk Ratio and Its Practical Implications

Marco Piccininni¹, Mats J. Stensrud²

¹Digital Health - Machine Learning Research Group, Hasso Plattner Institute for Digital Engineering, Potsdam, Germany

²Chair of Biostatistics, Institute of Mathematics, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Risk ratios are widely used effect measures in empirical research, but their stability and transportability across populations remain debated. Here, we show that the causal risk ratio is stable under selection based on immune status. For example, the causal risk ratio remains unchanged when individuals who cannot experience the outcome, regardless of treatment, are excluded from a study. We term this property "immune-selection stability" (ISS).

ISS applies broadly and generalizes previous findings on the stability of risk ratios. Furthermore, unlike earlier results, ISS does not rely on assumptions about cross-world counterfactuals. We also demonstrate an analogous property for survival ratios.

Despite decades of discussion on the properties of risk ratios, ISS has received little to no attention. However, its implications for interpreting, comparing, and transporting estimates across populations are considerable. We illustrate the practical relevance of ISS by discussing the results of hypothetical HIV trials.

4: Introducing an Open Access Simulated Benchmarking Data Resource to Enable Assessment and Neutral Comparison of Causal Inference Methods

Ruth Keogh¹, Nan van Geloven², Daniala Weir³

¹London School of Hygiene and Tropical Medicine (LSHTM), United Kingdom

²Leiden University Medical Center, Leiden, The Netherlands

³Utrecht University, Utrecht, The Netherlands

Observational data, such as from electronic health records, provide opportunities to address questions about causal effects of interventions under certain assumptions, and there is an extensive and growing literature on causal inference methods that enable this, including methods that combine statistical and machine learning techniques. The increasing availability and complexity of causal inference methods raises challenges for researchers making decisions about which methods to use in practice. Firstly, it is important to make comparisons of methods and their suitability for addressing different causal questions, but there is a lack of detailed and neutral comparisons, particularly using the type of observational data commonly faced in practice. Secondly, there is a lack of openly accessible data that researchers can use to learn, and teach others, how to implement methods and assess their practical feasibility.

In this work we have developed a simulated data resource designed to mimic complex longitudinal observational data. The data resource is intended to enable comparisons of methods for addressing a range of different types of causal question, as well as helping researchers to learn new methods. The data is based on a case-study concerning choice of second-line treatments for people with type-2 diabetes. It mimics longitudinal data including time-dependent treatments, longitudinal covariates of different types, and time-to-event outcomes, including competing events.

The data resource will be introduced and its potential use as a benchmarking data set, a template for a simulation study, or an educational tool will be discussed. Benchmarking data sets are real data sets to which different methods can be applied and compared, but they have the disadvantage that the data generating mechanism is unknown. Simulated data sets have the advantage of a known data generating mechanism, but they tend to be simpler than real data and have been criticised for being designed to favour some analysis methods over others. Our simulated data resource combines some of the benefits of real and simulated data, by mimicking the complexities of real data, while retaining the advantage that the data-generating process is known. The data-generating mechanism is designed so as not to favour particular analysis methods, therefore enabling more neutral comparison studies.

The use of the data will be illustrated with a comparison of causal inference methods for estimating treatment effects on time-to-event outcomes, including g-methods and doubly-robust

methods incorporating machine learning techniques. The potential for further development of the resource by the community will also be discussed.

5: Simulating Data from Marginal Structural Models for a Survival Time Outcome

Shaun Seaman¹, Ruth Keogh²

¹University of Cambridge, United Kingdom

²London School of Hygiene and Tropical Medicine

Marginal structural models (MSMs) are often used to estimate causal effects of treatments on survival time outcomes from observational data when time-dependent confounding may be present. They can be fitted using, e.g., inverse probability of treatment weighting (IPTW). It is important to evaluate the performance of statistical methods in different scenarios, and simulation studies are a key tool for such evaluations. In such simulation studies, it is common to generate data in such a way that the model of interest is correctly specified, but this is not always straightforward when the model of interest is for potential outcomes, as is an MSM. Methods have been proposed for simulating from MSMs for a survival outcome, but these methods impose restrictions on the data-generating mechanism. Here we propose a method that overcomes these restrictions. The MSM can be, for example, a marginal structural logistic model for a discrete survival time or a Cox or additive hazards MSM for a continuous survival time. The hazard of the potential survival time can be conditional on baseline covariates, and the treatment variable can be discrete or continuous. We illustrate the use of the proposed simulation algorithm by carrying out a brief simulation study. This study compares the coverage of confidence intervals calculated in two different ways for causal effect estimates obtained by fitting an MSM via IPTW.

Innovation in Oncology Dose Escalation Trials and Beyond

Wednesday, 2025-08-27 14:00 - 15:30, ETH E27

Chair: Sebastian Weber

1: Declaring Doses as Safe in Ongoing Oncology Dose-Escalation Trials: Are They Truly Safe? a Critical Assessment of Safety Criteria

Stefan Englert¹, Thomas J. Prior², Anirban Mitra³, Busola Sanusi², Liangcai Zhang², Illa Di Scala⁴

¹Janssen-Cilag GmbH, a Johnson & Johnson company, Germany

²Janssen Research & Development, LLC, a Johnson & Johnson company, USA

³Johnson & Johnson Limited, India

⁴Actelion Pharmaceuticals Ltd, a Johnson & Johnson Company, Switzerland.

Background / Introduction Ongoing dose-escalation trials present unique challenges in assessing safety, all with the goal to establish the maximum tolerated dose (MTD) and/or the recommended phase II dose (RP2D). Due to the extended duration of trials, it is common to declare doses as safe even if the entire adaptive and iterative dose escalation process has not been completed. This practice is often applied to aid trial management decisions, such as backfilling previous cohorts, allowing intra-patient dose escalation, or initiating concurrently combination therapies, to enhance efficiency and accelerate the pace of drug development.

There exists a misperception within the research community that once an escalation assessment team has cleared a cohort for escalation, the associated dose level is inherently safe. Zhao et al. (2024) referred to this as being *cleared for safety*. Given that dose escalation algorithms permit repeated de-escalations, previously cleared doses may not be truly safe, necessitating a quantification of this risk.

Methods We investigated criteria for declaring doses as safe, focusing on the BOIN design. Through rigorous simulation studies, we identified criteria that must be met to confidently declare dose levels as safe, minimizing the risk of treating additional patients at doses that exceed the MTD.

Results Our findings suggest that the criteria proposed by Zhao et al. are overly permissive, with up to 12.4% of doses classified as safe being above the MTD as established at the trial's conclusion. By mandating clearance of two consecutive dose levels, this percentage decreases to less than 1%. A more practical approach requires a minimum of three subjects

at the next dose level, which keeps the misclassification rate below 5% while allowing, on average, 18% more doses to be deemed safe.

Conclusions Criteria for declaring doses as safe are often neither specified nor are the implications of different safety rules examined. This may put patients at risk of receiving doses above the MTD determined only at the conclusion of the trial.

Our comparative analysis of safety criteria shows that evaluating a minimum of three subjects at the dose level above the one to be declared safe is essential for accurate safety assessments and protecting patients from toxic therapies. More aggressive dose-escalation paradigms would risk exposing patients to doses that are not ultimately declared safe.

References Zhao Y et al., "Backfilling Patients in Phase I Dose-Escalation Trials Using Bayesian Optimal Interval Design (BOIN)," *Clinical Cancer Research*, 30(4), 673-79, 2024.

2: Guiding Phase I Dose Escalation for Modern Oncology Therapies: Tackling (Informative) Dropout with Bayesian Multi-Cycle Time-to-Event Models

Lukas Andreas Widmer, Sebastian Weber

Novartis Pharma AG, Basel, Switzerland

The design of phase I trials in oncology predominantly utilizes a dose-escalation approach, monitoring dose-limiting toxicity (DLT) during the first treatment cycle to systematically determine the maximum tolerated dose (MTD). While appropriate for cytotoxic treatments, with the advent of targeted therapies, immunotherapies, and chimeric antigen receptor T-cells, this traditional model is increasingly inadequate for modern oncology treatments.

For contemporary non-cytotoxic therapies, efficacy and safety mechanisms are no longer inherently linked, rendering the "target toxicity" concept to establish a therapeutic dose as obsolete. Consequently, the MTD may not represent the optimal dose for further development, as the primary objective is to maximize efficacy while ensuring sufficient tolerability within a robust safety framework.

Furthermore, long-term administration requires sustained tolerability and efficient patient enrollment by leveraging partially-observed data. Long-term tolerability assessments and treatments for aggressively progressive cancers necessitate addressing patient dropout scenarios, which often lead to informative censoring. Aggressive cancers can cause dropout due to lack of efficacy at lower doses, while higher doses might result in dropout due to tolera-

bility issues such as sustained lower-grade adverse events. Traditional dose-toxicity models typically overlook these dropout scenarios.

Additionally, appropriate dosing regimens are crucial to prevent serious adverse events. For instance, T-cell engaging therapies might induce cytokine release syndrome post-treatment initiation, which can be mitigated through an incremental within-patient dosing strategy.

To address these complexities, we propose a Bayesian multi-cycle time-to-event (TTE) safety model, which extends existing frameworks such as the Bayesian Logistic Regression Model (BLRM) and Escalation With Overdose Control (EWOC). This TTE model efficiently leverages trial data to predict adverse event rates at various time points across multi-cycle therapies, beyond just the initial cycle. It also facilitates rapid cohort enrollment by accommodating partially-observed data due to ongoing enrollment or dropout.

We present considerations for developing these models through concrete case studies, illustrating model structures, priors, key data scenarios, and selected operating characteristics. Particular emphasis is placed on the impact of informative dropout on model performance, accurate MTD determination, patient safety during ongoing trials, and efficiency in terms of trial duration. Our findings demonstrate that the proposed TTE model serves as a viable alternative to the BLMR, particularly under conditions of patient dropout. Last, but not least, we hint at how these TTE models can further be extended to consider between-patient heterogeneity by including partially available pharmacokinetics data.

3: Potential Responses and the Order of Patient Inclusion in Early-Phase Sequential Trials.

Meliha Akouba¹, Matthieu Clertant¹, Alexia Iasonos², John O'quigley³

¹Université Sorbonne Paris Nord, France

²Memorial Sloan-Kettering Cancer Center, U.S.A

³University College London, U.K

Phase I clinical trials are the first tests of a new agent on human patients. The primary objective during this phase is to determine the maximum tolerated dose (MTD). Various statistically guided designs are used to sequentially allocate patients across dose levels. Traditionally, each dose is associated with a toxicity rate, defined by the proportion of patients experiencing dose-limiting toxicity (DLT). This conventional approach assumes a large, homogeneous patient population, summarized by an increasing vector of toxicity probability β , but overlooks the heterogeneity of the actual trial population, leading to potential issues in

accuracy and reproducibility. To address this limitation, we treat the trial population as fixed, meaning we condition on the specific patients enrolled in the trial. Each patient's potential responses are represented as a vector of binary treatment effects across dose levels, assuming monotonicity assumption. The overall toxicity rate in the trial population is then summarized by an increasing vector R . By eliminating random variability, this approach ensures full reproducibility of results, unlike analyses based on superpopulations (β). Furthermore, variance decomposition allows us to assess the impact of patient inclusion order, enabling a more precise evaluation of method performance. Our experiments indicate that, while traditional analyses might suggest only minor differences between methods, significant quantitative differences emerge when using R . We show both numerically and theoretically that comparisons based on the superpopulation β introduce nuisance variability, which tends to obscure competitive differences. For example, the mean squared error (MSE) difference between two designs might appear to be 15% under Beta but rises to approximately 25% under R , highlighting more pronounced disparities. Our methodology provides a comprehensive assessment of interval-based designs, including the Bayesian Optimal Interval (BOIN) method, which is the only Phase I design officially recognized as fit-for-purpose by the FDA and is now widely used in early-phase oncology trials. Our findings reveal that BOIN underperforms compared to dose-response model-based designs and even the historical "3+3" design. Moreover, its dose exclusion rule—intended to prevent the reassessment of highly toxic doses—exhibits considerable variability depending on the order of patient inclusion. This instability results in the exclusion of the true MTD in approximately one in five trials with a standard sample size of 36 patients, ultimately increasing the probability of recommending suboptimal doses in terms of efficacy.

4: How Best to Allocate Backfill Patients in Dose-Finding Oncology Trials: a Methodological Review & Simulation Studies Assessing Best Performance

Elli Bourmpaki¹, Helen Barnett², Oliver Boix³, Hakim-Moulay Dehbi¹

¹Comprehensive Clinical Trials Unit, University College London, United Kingdom

²School of Mathematical Sciences, Lancaster University, United Kingdom

³Bayer AG, Leverkusen, Germany

Background Dose-finding oncology trials (DFOTs) are a crucial step in research detecting the optimum dose of potentially effective anticancer therapies. Patients' allocation in DFOTs vary depending on the trial designs used. Novel approaches have been adopted recently by including additional patients at lower doses, these patients are referred to as backfill patients. Several statistical methodologies have been developed considering different ways of allocating backfill patients, however it is not clear which allocation approaches are better than others and

under which clinical scenarios. Therefore, a methodological review and further exploration on the impact of various allocation schemes for backfill patients is required.

Methods We will conduct a literature review using MEDLINE to explore the uptake of various allocation schemes for backfill patients in published DFOTs' results and developed methodologies between 2014 to 2024. Our aim is to assess: i) how many methods are developed for allocating backfill patients, ii) how many trials have used backfilling, iii) what allocation schemes have been used for backfill patients, iv) what trial designs were used, v) how was the recommended phase 2 dose selected using dose-finding and backfill patients' data. All possible ways of allocating backfill patients will be reported based on the developed methods and published trial results. The performance of selected allocation schemes will be explored using a simulation study. Patient and public involvement and engagement have contributed in this research by discussing potential allocation schemes for backfill patients.

Results For all eligible DFOTs we will report the proportion of various allocation schemes used for backfill patients, trial designs used, and developed methods for selecting the recommended phase 2 dose.

Conclusions Our review will show how widespread is the use of backfilling in DFOTs and what approaches are currently used to allocate backfill patients. Following our review, we will list all possible allocation schemes and use simulation studies to investigate the performance of selected schemes under specific trial designs and clinical scenarios. This body of work may be used to shape future trial design, conduct and analysis guidance for DFOTs considering backfilling.

5: Dose Optimisation in Early Phase Oncology Trials - Backfill and Expansion Cohorts

James Willard¹, Thomas Jaki^{1,2}, Burak Kürsad Günhan³, Christina Habermehl³, Anja Victor³, Pavel Mozgunov¹

¹MRC Biostatistics Unit, University of Cambridge, Cambridge, United Kingdom

²Faculty of Informatics and Data Science, University of Regensburg, Regensburg, Germany

³Merck Healthcare KGaA, Darmstadt, Germany

Historically, early phase dose finding trials in oncology focused on identifying the correct doses of cytotoxic chemotherapies, where more benefit was expected from higher doses. Therefore, these trials were traditionally designed to find a maximally tolerated dose (MTD), defined as the largest dose which satisfies specific toxicity constraints. Recently, the FDA's Project

Abstracts of Contributed Talks

Optimus highlighted how modern targeted therapies can provide benefit at doses lower than the MTD and so identifying these doses via dose optimisation has become a major objective of early phase trials. Unfortunately, the small sample sizes and short observation periods of these trials make dose optimisation challenging, since it is difficult to collect comprehensive information on the dose response curves under these settings. This may result in suboptimal doses being recommended for future development, adversely affecting patients and all later phase studies. To help remedy this and collect more information on the dose response curves before recommending doses for further study, the use of backfilling and expansion cohorts has been proposed. During dose escalation, backfilling cohorts are assigned to doses lower than the current estimate of the MTD. After dose escalation, expansion cohorts are assigned to a small number of the most promising doses. In this work, we examine the relationship between the escalation and expansion components of dose optimisation. We compare the performance of a variety of adaptive stopping rules which determine when to terminate escalation and transition to expansion. Furthermore, we investigate how the timing and number of patients used for backfilling may impact the selection of the doses used in expansion. Findings from an extensive simulation study will be discussed and recommendations for performing dose optimisation with backfilling and expansion cohorts will be provided.

Missing Data and Imputation

Wednesday, 2025-08-27 14:00 - 15:30, ETH E23

Chair: Jonathan Bartlett

1: Missing Value Imputation Methods in Prediction Model Development: a Neutral Comparison of Approaches

Manja Deforth^{1,2}, Georg Heinze³, Ulrike Held¹

¹Department of Biostatistics at the Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Zurich, Switzerland

²MSD Merck Sharp & Dohme AG, Zurich, Switzerland

³Center for Medical Data Science, Institute of Clinical Biometrics, Medical University of Vienna, Vienna, Austria

Background In the development of prognostic models, missing predictor data are quite common and can be handled by using imputation methods. In a neutral comparison study, we aimed at comparing three popular imputation methods by simulating data resembling a Swiss multicenter prospective cohort study on long COVID.

Methods Assuming a binary outcome and nine predictors, we designed 36 scenarios, resulting from different sample sizes, proportions of noncomplete cases and missingness mechanisms in 3 predictors. The missing data was imputed by applying three imputation algorithms: missForest, aregImpute and mice. The missForest algorithm is based on a random forest methodology, while aregImpute uses flexible additive imputation models and samples drawn from a predictive posterior distribution. In mice linear-additive models are used for imputations. We conducted a single imputation for missForest, 5 and 100 imputations for mice and 100 imputations for aregImpute. Prediction models were estimated on the imputed datasets using linear-additive logistic regression. We also performed complete case analysis without imputing the missing data. All prognostic models were validated on validation cohorts without missing values by evaluating overall performance (scaled Brier score), model discrimination (*c*-statistic), calibration intercept and slope.

Results Complete case analysis resulted in the lowest prediction model performance, and for the imputation methods a higher proportion of missing values was associated with lower model performance. Scaled Brier scores from mice and aregImpute models were higher than for missForest. aregImpute preformed remarkably well, yielding a calibration slope close to

one which was even higher than if no data were missing. Model calibration was influenced more strongly by the imputation method than model discrimination.

Conclusion Using aregImpute for the imputation of missing values resulted in shrinkage of regression coefficients of the prediction model, leading to a near optimal calibration slope. The usage of multiple imputed methods such as mice and aregImpute can be recommended in most cases, and when it is not possible to ascertain the true values of missing data.

Publication: Deforth M, Heinze G, Held U. The performance of prognostic models depended on the choice of missing value imputation algorithm: a simulation study. *J Clin Epidemiol* 2024; 176:111539.

Disclaimer: This work was done while the first author was working at the University of Zurich.

2: Assessing the Impact of Percentage of Missing Data and Imputation Methods on Youden Index Estimation

Sergio Sabroso-Lasa¹, Luis Mariano Esteban², Tomás Alcalá-Nalvaiz³, Núria Malats¹

¹Genetic and Molecular Epidemiology Group, Spanish National Cancer Research Centre (CNIO) and CIBERONC, Madrid, Spain

²Department of Applied Mathematics, Escuela Universitaria Politécnica La Almunia, University of Zaragoza, Zaragoza, Spain

³Department of Statistical Methods, University of Zaragoza, Zaragoza, Spain

Background

The rapid advancement of computational methods and data collection technologies has resulted in an exponential increase in the generation of large, complex datasets with numerous variables, often sourced from multiple origins. As a result, managing missing data has become a critical challenge in statistical modeling, especially in large databases, where the percentage and distribution of missing data can significantly affect analytical outcomes.

Effectively imputing missing values is crucial for ensuring the reliability of statistical inferences and predictive models. Although various imputation techniques are available, understanding how the proportion and distribution of missing data influence model performance remains a complex issue that requires further investigation. While previous research has examined the impact of missing data on model discrimination through metrics like the area under the ROC

curve (AUC), the specific effect on the Youden Index (J), a key measure of test effectiveness that combines sensitivity and specificity, has not yet been explored.

Method(s) and Results

We conducted simulations under realistic conditions to assess the impact of missing data on the estimation of the Youden Index. These scenarios included independent normally distributed variables with varying predictive capacities, predefined correlation structures, categorical variables, and skewed distributions. Additionally, we analyzed cases where missing data followed specific predefined patterns.

We applied various imputation methods, including MissForest, Multivariate Imputation by Chained Equations (MICE), and K-nearest neighbors, to evaluate the predictive value of the models across different levels of missing data, which ranged from 5% to 75%. The effectiveness of each method was assessed using key diagnostic metrics such as AUCs, sensitivity, specificity, and the Youden Index.

Our findings indicate that most diagnostic metrics decrease by 20–30% compared to models with complete data, except for specificity, which remains comparatively robust. Importantly, the Youden Index varies significantly based on the proportion of missing data, highlighting the challenge of establishing an optimal cutoff point in clinical practice when dealing with incomplete datasets.

Conclusions Our findings underscore the significant impact of missing data on diagnostic metrics, particularly regarding the proportion of missing values. While most predictive models incorporate an imputation step, few account for how the distribution of missing data influences overall model performance. This analytical oversight restricts potential enhancements in evaluation metrics and can lead to unreliable Youden Index values and cutoff points. These findings highlight the necessity for further research to refine imputation strategies and improve the reliability of predictive models in the context of missing data.

3: Recorded Reasons of Missingness-Informed Sensitivity Analyses in Clinical Trials

Dries Reynders¹, Jammbe Musoro², Saskia le Cessie³, Els Goetghebeur¹

¹Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium

²European Organisation for Research and Treatment of Cancer (EORTC) Headquarters,

Brussels, Belgium

³Department of Biomedical Data Sciences and Department of Clinical Epidemiology, Leiden University Medical Center, The Netherlands

In clinical trials, missing longitudinal outcome data, like patient reported outcomes (PROs) is most often addressed assuming missing at random (MAR) and fitting a mixed model, sometimes accompanied by missing not at random (MNAR) sensitivity analyses. Typically, these sensitivity analyses add a shift towards worse outcomes to the MAR-imputed values or use reference-based imputation as a one size fits all approach.

This practice leaves room for more tailored approaches where different categories of missing outcomes are handled differently. If available, categorizing can be based on recorded reasons for missingness (too ill, inconvenient, administrative failure,...). These allow not only to assess the credibility of the MAR-assumption but also may inform more differentiated and realistic MNAR sensitivity-scenarios. To this end, we can link the recorded reasons for missingness to later observed outcomes.

For intermittent missingness, surrounding observations may display very different patterns across reasons for missingness. If these patterns essentially reflect the underlying truth – e.g. observed stable patterns in patients with missingness due to administrative failure, but tendency to decline in patients who were ill at some visit – assuming MAR may still be reasonable. On the other hand these differences may inform different MNAR-imputation scenario with handling missingness dependent on the reason.

With attrition, the longitudinal outcomes are no longer observed, but adverse events and intercurrent events like disease progression or death may still be recorded. These too present a vehicle to set up more credible sensitivity analyses or MAR-imputation.

The usefulness of the information captured by the recorded reasons for missingness, is not confined to the longitudinal analysis itself. If reasons for missingness prove to be predictive of death or drop-out, it may also be used to relax the non-informative censoring assumption for the time-to-event analyses. Incorporating these reasons in inverse probability of censoring weighting-analyses may then provide more robust evidence.

In a large randomized oncology trial with high mortality comparing radiotherapy alone with adjuvant and concomitant chemotherapy, we investigate in depth how recorded reasons for missing patient reported outcome-data (PRO) are linked to observed outcomes, overall survival and its censoring. Building on the resulting insights, we set up sensitivity analyses for the PRO's and overall survival. Comparing these with more standard analyses, reveals the value of recording these reasons.

4: Adjusting for Outcome Reporting Bias in Meta-Analysis: a Multiple Imputation Approach

Cora Burgwinkel¹, Leonhard Held^{1,2}

¹Epidemiology, Biostatistics and Prevention Institute (EBPI), Universität Zürich, Switzerland

²Center for Reproducible Science (CRS), Universität Zürich, Switzerland

Background Outcome reporting bias (ORB) occurs when research study outcomes are selectively reported based on their results. ORB potentially undermines the credibility and validity of meta-analyses and contributes to research waste by distorting overall treatment effects. ORB can be viewed as a missing data problem where unreported outcomes introduce bias. Despite the serious implications ORB poses, it remains an underrecognized issue, with only a few adjustment methods available.

Methods We propose an approach that addresses unreported studies in meta-analyses through multiple imputation. The imputed data are reweighted using importance sampling to provide an adjusted estimate of the treatment effect, building on existing methods for selection bias from the literature [1]. To assess the impact of ORB in meta-analyses of clinical trials, we apply our proposed methodology to real clinical data affected by ORB. Additionally, we conduct a simulation study to evaluate the method's performance, focusing on treatment effect estimation across varying degrees of selective non-reporting.

Results The proposed method successfully adjusts for ORB under assumptions of selective non-reporting. The results demonstrate that ORB can significantly affect the conclusions of a meta-analysis, particularly when the number of unreported studies is large.

Conclusion Imputing unreported outcomes provides a promising approach to address ORB in meta-analyses. The method assumes a specific mechanism for non-reporting and has been applied exclusively to summary-level data. Besides, so far only the univariate approach has been explored, meaning ORB adjustment was investigated separately for each outcome. Further research is required to extend our approach to multivariate meta-analysis, allowing for simultaneous adjustment of multiple outcomes. Additionally, applying the proposed method on individual patient data (IPD) could provide more precise and reliable ORB adjustment.

References

- [1] James Carpenter, Gerta Rücker, and Guido Schwarzer. Assessing the sensitivity of meta-analysis to selection bias: a multiple imputation approach. *Biometrics*, 67(3):1066–1072, 2011.

5: Comparing Estimation Methods for Expected Quality of Life and Predicted Patient-Specific Trajectories in Oncology: Addressing Missingness/Death (not) at Random

Eline Vanderpijpen, Els Goetghebeur

Ghent University (Belgium)

Cancer treatments affect both patient survival and quality of life (QoL). Insights into expected QoL and the variation in patient-specific QoL predicted under different treatments may therefore support treatment selection when combined with survival curves. The survival curve with expected ‘QoL measures - while alive’ nevertheless remains an important two-dimensional target estimand for treatment policy evaluation.

We adapted and compared various statistical methods to assess average QoL - while alive and predict individual QoL trajectories under different treatments. Missing data following intercurrent events, like treatment discontinuation, are common in this setting, and at least missing at random (MAR), while death is likely not at random (DNAR). Weighted generalized estimating equations (WGEE), mixed models, and joint models allow in their own way for MAR missingness in the longitudinal setting besides DNAR.

At specific time points the average QoL among the living can be estimated using weighted GEE - or double robust estimators - weighting for missingness but not death at each time point. This performed well, even with data simulated under joint models. In contrast, standard mixed and joint models naturally target a hypothetical estimand, averaging QoL over an ‘immortal population’ by implicitly imputing QoL after death. When death is not at random, such mixed model estimates are biased for this hypothetical estimand. By using a shared parameter model framework to analyse both survival data and QoL measurements, joint models avoid this bias. We found their derived parametric estimators for the average QoL *among the living*, to be less accurate than those from WGEE, possibly due to finite samples.

Mixed models and joint models can predict patient-specific QoL trajectories once the random effects are estimated from best linear unbiased predictors, or expectation-maximization (EM) estimators, respectively. Under DNAR, predictions from mixed models were again biased, but joint model predictions performed reasonably. With normal random effects, however, EM-estimates shrunk towards the mean QoL, generating overly optimistic predictions for the most ill patients. For improved accuracy, we explore alternative methods for estimating random effects, including a mixture of prior normal distributions.

We perform a phase 1 simulation study and re-analysed a randomized oncology trial with

Abstracts of Contributed Talks

high mortality rates using the methods described. Our findings support using WGEE for the estimands, while reworked joint models may allow to evaluate QoL in cancer trials for more personalized treatment decisions.

Observational/Real-World Data 1

Wednesday, 2025-08-27 14:00 - 15:30, ETH E21

Chair: Michail Katsoulis

1: Handling Informative Patient Monitoring in Routinely-Collected Data Used to Estimate Treatment Effects, with Application to High-Frequency Hospital Data

Leah Pirondini¹, Karla Diaz-Ordaz², Ruth Keogh¹

¹Department of Medical Statistics, London School of Hygiene and Tropical Medicine, UK

²Department of Statistical Science, University College London, UK

Introduction and Objectives Routinely-collected hospital data provide opportunities to gain understanding of treatment effects that would not be feasible in randomised trials and that reflect their impact in realistic clinical practice. A challenge presented by hospital data is that measurements of patients' clinical status are made at high frequency, on differing schedules for each patient dependent on their underlying clinical status, so timing and frequency of measurements is informative. However, many existing causal inference methods assume measurements are made at regular time intervals. The aim of this work is to evaluate methods for estimating causal effects of longitudinal treatments in the presence of informative monitoring. This is motivated by hospital data on patients in the intensive care unit and questions about optimal mechanical ventilation strategies.

Methods and Results We compare methods based on (i) marginal structural models fitted by inverse probability of treatment weighting (MSM-IPW), (ii) G-computation, and (iii) longitudinal targeted maximum likelihood estimation (LTMLE). We assume an underlying grid of time, such that time-dependent variables are either monitored or unmonitored at each time-point. Methods are based either on imputation of unmonitored covariate data or on adapting inverse probability weights to account for monitoring variables. We evaluate methods using a simulation study, comparing against more simple approaches using last-observation-carried-forward (LOCF) ignoring informativeness of monitoring. Data are simulated to represent a range of realistic scenarios with time-varying treatment and covariates, in which monitoring depends on past covariate, treatment and monitoring levels. We also illustrate methods in a real-world example using routinely-collected intensive care data from UCLH to investigate the use and the timing of initiation of invasive mechanical ventilation vs non-invasive or no ventilation on mortality.

We show that ignoring monitoring can result in bias, the size of which depends on infor-

mativeness of the monitoring process. All methods reduce bias compared with their naïve LOCF-based equivalents, with LTMLE and G-computation based methods resulting in the smallest bias.

Conclusions Data with informative monitoring are common in observational studies, but there is a lack of readily-implementable methods to handle them. We describe three methods and evaluate their performance.

2: Target Trial Emulation to Duplicate Randomized Clinical Trials using Registry Data in Multiple Sclerosis

Antoine Gavoille^{1,2,3}, Mikail Nourredine^{2,3}, Fabien Rollot⁴, Romain Casey⁴, Sandra Vukusic^{1,4}, Muriel Rabilloud^{2,3}, Fabien Subtil^{2,3}

¹Hospices Civils de Lyon, Service de Neurologie, sclérose en plaques, pathologies de la myéline et neuro-inflammation, F-69677 Bron, France

²Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Évolutive UMR 5558, F-69100 Villeurbanne, France

³Service de Biostatistique-Bioinformatique, Hospices Civils de Lyon, F-69003 Lyon, France

⁴Observatoire Français de la Sclérose en Plaques, Centre de Recherche en Neurosciences de Lyon, INSERM 1028 et CNRS UMR 5292, F-69003 Lyon, France

Introduction Target trial emulation (TTE) offers a rigorous framework to answer causal questions using observational data and could be of major interest to the field of multiple sclerosis (MS) research. Replicating the results of randomized clinical trials (RCTs) is a key approach to validate the TTE methodology and the data source used. In the present study, we aimed to replicate 8 RCTs evaluating the efficacy of an active disease-modifying therapy (DMT) versus a treated control group in MS using observational data from the French MS registry, and to compare different g-methods.

Method Data were extracted in December 2023 from the *Observatoire Français de la Sclérose en Plaques* (OFSEP) database. For each emulated trial, we included patients who initiated one of the DMTs evaluated in the trial and met its inclusion criteria, and compared the initiation of the active DMT vs. control DMT, in an intention-to-treat setting. The primary outcome was the annualized relapse rate (ARR). Secondary outcomes were EDSS progression confirmed at 3 months during the study period, and new/enlarged T2-lesions and new gadolinium-enhanced T1-lesions on a brain MRI during the study period. Several g-methods were applied to estimate the treatment effect adjusted for confounding factors between groups and corrected for censoring and missing outcome assessment: propensity-

matching, inverse probability weighting (IPW), g-computation, and targeted maximum likelihood estimator (TMLE). The concordance between the treatment effects estimated in emulated trials and in the corresponding RCT was analyzed using predefined agreement metrics.

Results A total of 14 111 patients were included in the 8 emulated trials: ASSESS, BEYOND, CONFIRM, OPERA, REGARD, RIFUND-MS, TENERE, and TRANSFORMS. Emulated trials estimates were concordant with RCT results in 7 of 8 trials for relapse rate, and in all 6 trials which evaluated EDSS progression. Radiological outcomes were more challenging to replicate, achieving concordance in 3 of 5 trials for the analysis of new T2-lesions, and 1 of 4 trials for new gadolinium-enhanced T1-lesions. Among g-methods, TMLE provided estimates most consistent with RCTs, while IPW and g-computation yielded comparable results but diverged in trials with fewer patients. Matching-based estimates showed higher variance and greater deviation from TMLE in smaller sample sizes.

Conclusion The use of a TTE methodology applied to the OFSEP registry data is a valid and powerful tool for evaluating treatment effectiveness in MS. Our results support the use of real-world evidence to explore questions beyond the scope of RCTs.

3: Exploring Synthetic Control Data Quality Between Data Types in Two Case Studies: COVID-19 and Crohn's Disease

Nicole Ann Cizauskas, Svetlana Cherlin, James Wason

Newcastle University, United Kingdom

Introduction Synthetic control arms are useful in clinical trials that have restricted numbers of participants available, such as for rare diseases. The current literature on creating synthetic controls suggests that randomised control trial (RCT) data is the best data source compared to observational study data or external data. This paper aims to provide a method for measuring and comparing the quality of synthetic control data, using two metrics: treatment effect maintenance and standard mean difference (SMD) between data types.

Methods Two case study datasets were selected to illustrate this, COVID-19 and Crohn's Disease. For each case study, RCT data, observational study data, and external real-world data were selected and compared. Datasets were simulated from summary level data in real studies, and synthetic data was produced from these simulated datasets. Four scenarios with differing sample sizes were simulated to test the effect of sample size on synthesis quality. Three different data synthesis methods were compared: categorical and regression tree (CART) models, linear/logistic regression, and random sampling. The treatment effect

on the disease outcome was measured using a chi-squared test. SMD was calculated between each simulated variable and its corresponding synthetic variable in each dataset. SMD was also calculated between corresponding variables of different data types (RCT, observational, and external) and compared across both simulated and synthetic datasets.

Results The metrics show little difference in quality between RCTs and other data types in the two disease case studies tested. There were no notable differences between sample size scenarios or method of data synthesis in either treatment effect maintenance or SMD. Quality did fluctuate across synthetic datasets, but not in an identifiable pattern.

Discussion Future studies looking to use synthetic controls should not disregard the use of observational study or external data in the creation of synthetic controls but should check the quality of any synthetic control groups created regardless. Testing this method on other disease datasets would provide a better understanding of how data type influences synthetic data quality.

4: The Most Appropriate Method for Outlier Detection in a Clinical Audit Depends on the Data Distribution

Anqi Sui, Menelaos Pavlou, Rumana Z. Omar, Gareth Ambler

University College London, United Kingdom

Introduction Monitoring the clinical performance of healthcare units (e.g. hospitals, surgeons) is essential for national audits, particularly in identifying 'outlier' units whose performance (e.g. probability of in-hospital death) deviates significantly from expected performance. Detecting and managing outliers is crucial for improving healthcare quality.

Common methods for outlier detection include Common Mean Model (CMM) and Random Effects Logistic Regression (RELR). Our study evaluates their performance through simulation and provides recommendations for their appropriate use.

Methods CMM assumes that the probability of death is the same in all units, attributing any observed differences to random binomial variation. As the observed variability is often larger than expected (overdispersion), CMM is applied with an overdispersion correction. To detect outliers, test statistics are constructed based on differences between observed and expected unit mortality; these are assumed to follow a normal distribution for 'in-control' units. In contrast, RELR uses test statistics based on the estimated random effects which are on the logit scale and assumed to follow a normal distribution for 'in-control' units. Both

assumptions cannot hold simultaneously unless outcome prevalence is close to 0.5.

To assess the performance of these methods when their assumptions are violated, we simulated scenarios with varying numbers of units, unit sizes, outcome prevalences, and levels of variability between units. Two data-generating mechanisms (DGMs) were used, based on CMM and RELR respectively. The performance of each method was assessed focusing on the overall false positive rate (FPR) and the FPR for 'good' and 'bad' (low/high mortality) outliers separately.

Results Both methods appeared to work well, achieving the nominal overall FPR. However, the FPR for good and bad outliers deviated from the nominal level when the DGM was not aligned with the outlier detection method. When outcome prevalence was low, applying CMM to RELR-DGM data led to over-detection of bad outliers and under-detection of good outliers (and vice versa). These issues worsened by small unit sizes and greater variability between units. Both methods were applied to real datasets with low prevalence leading to differences that can be attributed to the findings above.

Conclusion CMM and RELR are widely used in clinical audits for outlier detection. Our findings reveal that violations of their underlying assumptions can have serious implications, potentially leading to unfair scrutiny of healthcare units or failing to flag underperforming units. The most appropriate method should be chosen following a check of the test statistics distribution, e.g. using appropriate diagnostic tools.

5: Incorporating Real-World Data to Refine the Calculation of Probability of Success

Bergas Fayyad^{1,2}, Laura Rodwell¹, Kit Roes¹, Giulia Ferrannini², Christian Basile², Lars Lund², Gianluigi Savarese², Aysun Cetinyurek-Yavuz¹

¹Radboud University Medical Center, Netherlands

²Karolinska Institutet, Sweden

Background In drug development, several trials are required to progress to confirmatory evaluation. A crucial milestone is the decision on whether to proceed to phase III based on phase II results. Several quantitative methods have been developed to inform this decision, with one of the more widely used being probability of success (PoS). In some cases, the endpoint used in phase II and phase III trials are different. Several approaches have been proposed to address this difference. However, a recent review [1] highlighted the potential

of using real-world data (RWD) for that purpose. We focus on the case where the phase II trial uses a continuous biomarker endpoint while the planned phase III trial uses a survival endpoint. We propose a method to construct the “design prior” for the primary survival endpoint which incorporates the association between the biomarker and the survival endpoint estimated from RWD.

Methods The association between the biomarker and the survival endpoint is first obtained through a Cox proportional hazard model using registry data. This association is then combined with the biomarker treatment effect estimate from phase II trial to obtain the treatment effect on the survival endpoint. This approach can also incorporate a prior distribution directly on the hazard ratio of the survival endpoint if the information is available (e.g., from phase II). We demonstrated this approach using the Swedish Heart Failure Registry. We compared the impact of important data-related decisions in using registry data, including the timing of the follow-up and biomarker measurement, choice of endpoint, and the relevant patient population.

Results and conclusion With a well-established registry, it was possible to derive estimates of PoS, including exploring for relevant subgroups. The settings in using registry data had an impact on the association between the biomarker and survival endpoints, and thus the PoS. The choice of the subset of patients had the largest impact of the registry-related aspects. The change in PoS due to the inclusion of a prior distribution directly on the hazard ratio was larger than any specifications related to the registry data. This work provides a methodological solution to incorporate registry data in the PoS calculations to aid decision-making.

References 1. Cetinyurek Yavuz, A., et al., *On the Concepts, Methods, and Use of “Probability of Success” for Drug Development Decision-Making: A Scoping Review*. Clinical Pharmacology & Therapeutics, 2025.

Efficient Use of Interim Analyses in Clinical Trials

Wednesday, 2025-08-27 16:00 - 17:30, Biozentrum U1.131

Chair: Francois Mercier

1: Innovative Clinical Trial Approach for Evaluating Digital Medical Devices under European Fast-Track Regulatory Frameworks

Moreno Ursino¹, Sandrine Boulet¹, Raphaël Porcher², Edouard Lhomme^{3,4,5}, Florence Francis-Oliviero^{3,4}, Gaël Varoquaux⁶, Florence Saillour^{3,4}, Corinne Collignon⁷, Rodolphe Thiébaut⁸, Sarah Zohar¹

¹Inserm, UMRS 1346, Université Paris Cité, Inria, HeKA, F-75015 Paris, France

²Université Paris Cité and Université Sorbonne Paris Nord, Inserm, INRAE, Center for Research in Epidemiology and Statistics (CRESS), F-75004 Paris, France

³Université de Bordeaux, ISPED, INSERM, Bordeaux Population Health Research Center, U1219, Bordeaux, F-33000, France

⁴Service d'Information médicale, CHU de Bordeaux, Bordeaux, F-33000, France

⁵INRIA SISTM team, Talence, France

⁶INRIA Soda team, Palaiseau, France

⁷Haute Autorité de Santé, Saint-Denis, France

⁸Université de Bordeaux, Inserm, Inria, Bordeaux France

Background / Introduction Recently, several health technology assessment bodies across European Union countries have begun incorporating users' demands for quicker access to digital medical devices (DMDs). Introducing a conditional fast-track pathway that enables early access and reimbursement presents an attractive solution for patients, healthcare professionals, and the medical device industry. However, regulators must ensure they have enough data to justify provisional reimbursement and user access, even if complete clinical evidence is not yet available. By the time manufacturers complete the clinical study, they will have also gathered real-world data (RWD) from the DMD's use in the population. This dual source of information—clinical trial data and RWD—provides regulators with a richer evidence base for final decision-making, a significant improvement over the traditional reliance solely on clinical trial data. We propose a statistical framework that integrates clinical trial data and RWD, allowing a more rigorous evaluation of DMDs in compliance with European fast-track regulatory requirements.

Methods Our framework includes three stages: (1) an interim analysis of clinical trial data to support temporary regulatory authorization and enable the collection of RWD; (2) a final

analysis of the clinical trial data; and (3) a meta-analysis that integrates RWD with clinical trial data, provided temporary authorization was granted. Various metrics are introduced to optimize the timing of the interim analysis and the application for temporary authorization. The framework was evaluated using a simulation study.

Results Using a significance level of 0.025 and a power of 0.9 for a one-sided two-sample test for proportions, (0.7, 0.4) as the true response rates for the intervention and control arms in the clinical trial population, $c^* = 0.8$ as the threshold for the conditional power (CP), the timing of the interim analysis is deduced as $t = 0.5$ and the CP is greater than c^* in 70% of cases. In these situations, the temporary regulatory authorization is obtained, and therefore RWD can be collected in parallel with the second part of the clinical trial. Then, integrating RWD into the final analysis allows a gain in effective sample size compared to the traditional approach.

Conclusion This framework should include a post-market evaluation of the DMD after its widespread adoption, in line with the principles of phase IV studies. While the primary purpose of the final augmented analysis is to refine the findings of clinical trials, it can also help to assess the generalizability of those results in real-world settings.

2: Using Only an Early Outcome for Interim Decisions Regarding Treatment Effect on a Long-Term Endpoint: a Practical Implementation

Tomasz Burzykowski^{1,2}, Leandro Garcia Barrado²

¹Hasselt University, Belgium

²IDDI, Belgium

In randomized clinical trials that use a long-term efficacy endpoint T , the follow-up time necessary to observe T may be substantial. This may limit the timing of interim analyses based on T . In such trials, an attractive option is to consider an interim analysis based solely on an early outcome S that could be used to expedite the evaluation of treatment's efficacy.

Garcia Barrado and Burzykowski (Pharmaceutical Statistics 2024) developed a methodology that allows introducing such an early interim analysis for any combination of S and T types. It appears that such a design may offer substantial gains in terms of both the expected trial duration and the expected sample size. A prerequisite, though, is that the treatment effect on S has to be strongly correlated with the treatment effect on T , i.e., the early outcome is a good trial-level surrogate for the long-term endpoint.

When developing their methodology, Garcia Barrado and Burzykowski assumed that the coefficients defining the (trial-level) model used to evaluate the properties of S as a surrogate for T were known. However, in practice, only estimates of the coefficients, obtained by using data from a meta-analysis, would be available. This fact importantly limits the applicability of the methodology. In the current manuscript, we address this issue.

Methods To adjust for the fact of estimation of the trial-level surrogacy model, the variance of the interim-analysis test-statistic has to be inflated by a term related to the variance-covariance of the estimated model coefficients. We obtain an explicit expression for the term by using measurement-error modelling. By applying the expression to a set of hypothetical scenarios for a clinical trial, we evaluate the gain in operating characteristics of a trial with an interim analysis based solely on data for S , with and without the adjustment for the estimation of the model. We also illustrate the application of the developed results by using a real-life clinical trial.

Results As expected, the adjustment leads to a reduction in the gain in operating characteristics. It appears, however, that if S is a good surrogate for T , the relative reduction may be small.

Conclusion The obtained results allow for designing trials with an interim analysis based only on an early outcome, while properly adjusting for the error resulting from the estimation of the model capturing properties of the outcome as a surrogate for the long-term endpoint.

3: Calibrated Risk-Scale: A Proposed Futility Design Framework to Enhance Portfolio-Level Profitability and Performance

Nima Shariati

F. Hoffmann-La Roche AG, Switzerland

Futility analysis is an effective adaptive design that enables trials to potentially be terminated at a predetermined interim stage. For pharmaceutical companies and clinical trial sponsors, balancing the risk of prematurely stopping a trial for a potentially successful drug, against the risk of continuing a trial when collected interim data suggests the drug is ineffective, is a complex challenge. The distinct nature, significance, financial and other implications of these risks make them too intricate to be easily aggregated for decision-making purposes. Furthermore, the inclusion, timing, and strictness of futility designs have a considerable spillover effect on the entire portfolio. The opportunity cost of making no (or suboptimal) interim futility gating decisions, which could free up financial and human resources for reinvestment

in other opportunities in the pipeline, exemplifies a portfolio-wide impact.

This work aims to introduce a refined framework to identify optimal futility designs for individual trials but viewed from a portfolio-level perspective. The framework seeks to balance the errors of falsely continuing and falsely stopping trials by weighing them according to their financial impact, both at the trial level and, more importantly, at the portfolio level.

Consequently, this framework quantitatively determines the extent of leniency and prudence in risk-taking based on the total portfolio-level financial impact of interim decisions while considering the mutual interconnectivity among various trials within the portfolio. Besides accounting for the unique financial characteristics of each trial, the proposed framework also evaluates the alternative uses of freed-up resources within the portfolio by assessing what could have been achieved if some trials had been prematurely stopped. This scheme can subsequently suggest suitable futility designs for each trial whilst ensuring overall portfolio-level optimization.

To demonstrate the sensitivity of the framework, several sensitivity analyses were conducted. These analyses primarily addressed the uncertainty surrounding the assumed drug effect at the trial's design stage, as well as the uncertainty related to cost evaluation caused by the potential loss of not investing in other opportunities within the portfolio due to the continuation of less effective and riskier trials.

4: Interim Analysis under Treatment Effect Heterogeneity

Audrey Boruvka

Hoffmann-La Roche Limited, Canada

Background In the conduct of interim analysis to adapt trial design, the U.S. FDA 2019 guidance "Adaptive Designs for Clinical Trials of Drugs and Biologics" emphasizes the need to control the risk of drawing erroneous conclusions and to reliably estimate the underlying treatment effect. Statistical methodology to achieve these objectives has long-been available to trial statisticians; however, they carry the essential assumption that the underlying treatment effect is homogeneous across the stages of the trial.

Methods Two settings in which homogeneity may become implausible are the presence of (1) underlying disease endotypes and (2) outcomes that depend on the patients' perception about their assigned treatment. Focusing on these scenarios and a variety of interim analysis objectives, we examine how knowledge about the heterogeneity may be incorporated into

either quantifying objective measures of risk or devising design adaptations.

Results With reasonably broad knowledge about the extent treatment effect heterogeneity, we derive an upper bound on error associated with decisions based on interim analysis for futility. When relatively little is known about the heterogeneity beyond its presence, we devise some general strategies to mitigate risk and stage effects and illustrate in the setting of design adaptations on multi-arm trials.

Conclusions Potential departures from treatment effect homogeneity must be carefully considered in planning any comparative interim analysis. Although heterogeneity may pose an insurmountable challenge for certain design adaptations, there are settings where sound decisions on the basis of interim analysis may still be made - particularly if one is willing to trade off benefits and costs for the sake of caution.

5: When to Schedule the Interim Analysis in the Presence of Missing Data?

Neža Dvoršak¹, Jianmei Wang², Thomas Burnett¹, Christopher Jennison¹, Robin Mitra³

¹University of Bath (United Kingdom)

²Roche (United Kingdom)

³UCL (United Kingdom)

Introduction Suppose an adaptive Phase III trial has an interim analysis scheduled at a given information fraction, e.g., 50%. The key question is: When will it reach 50% information? In a non-longitudinal setting, the information level for a continuous endpoint can be approximated by the fraction of patients with the endpoint data at the interim analysis relative to the final analysis. However, longitudinal trials with repeated measures and missing data require more nuanced methods to estimate the information level accurately. The question then becomes: When will there be 50% information in the presence of missing data? Is it when half of the patients reach the final visit, or could it be earlier?

Methods We propose an approach for projecting the information fraction at an interim analysis in a continuous longitudinal trial analysed using the MMRM framework. We establish a relationship between information time and calendar time, providing practical guidance. At the design stage, a prediction for the timing of interim analysis is based on assumptions about enrolment rate, total sample size, dropout rate, visit timing, and the correlation matrix between visits. Once some data are available, this prediction is refined using the observed enrolment rates, dropout patterns, and updated correlation estimates, yielding a more accurate

estimate of the current information level and an updated timeline for the interim analysis.

Results We demonstrate that we can project information timelines at the design stage and refine them as data accrues. In the context of a worked example, we show how to navigate different missing data patterns, assess the current information level, and set a reliable timeline for the interim analysis.

Conclusion Accurately estimating the timing of the interim analysis in longitudinal trials with missing data is essential for optimizing trial conduct, especially in terms of ethics and allocation of efforts and resources. Leveraging both initial design assumptions and accumulating trial data, our approach enhances decision-making, ensuring that interim analyses occur at the intended information fraction.

Survival Analysis 3

Wednesday, 2025-08-27 16:00 - 17:30, Biozentrum U1.141

Chair: Dominic Edmund Magirr

1: Estimating the New Event-Free Survival

Judith Vilsmeier¹, Maral Saadati², Kaya Miah², Axel Benner², Hartmut Döhner³, Jan Beyersmann¹

¹Institute of Statistics, Ulm University, Ulm, Germany

²Division of Biostatistics, German Cancer Research Center (DKFZ), Heidelberg, Germany

³Department of Internal Medicine III, Ulm University Hospital, Ulm, Germany

In leukemia studies, the endpoint event-free survival (EFS) is defined as time from diagnosis or study entry

until date of primary refractory disease, relapse or death, whichever occurs first. Since 2022 the European

LeukemiaNet (ELN) recommends for patients who are considered treatment failures, i.e. who are evaluable

for response but do not achieve remission by a pre-defined landmark or die before the landmark without

response assessments, to record the event at day 1. A similar recommendation holds for patients who are

alive but non-evaluable, only that they are censored at day 1. This leads to a potentially large drop of the

estimated EFS at day 1. However, the shift of some event times to day 1 has the consequence that the margin

of the drop is underestimated by the Kaplan-Meier estimator if patients are censored before the landmark.

Our aim is to present an unbiased estimate for EFS in which patients who are considered treatment failures

are accounted for in a way that is consistent with the intent of the recommendation. For this, "Event at day 1"

is defined as one event type and "Event after day 1" as a competing event and the Aalen-Johansen estimator

is used to estimate the event-specific transition probabilities, which are then combined in one EFS estimate.

In addition, we establish a formal link to cure models by equating the patients who are con-

sidered treatment failures with the "cured" proportion in cure model terminology and present inference methods.

2: Marginal Matched Pairs Cox Regression

Jana Kinzel, Jan Beyersmann

Institute of Statistics, Ulm University, Germany

The Cox regression model is an important method for analysing cohort survival data, with the consistency of its parameter estimator established under certain assumptions, such as the independence of all individuals. In practice, this assumption does not always hold. One such situation involves matched cohort datasets, where individuals are pairwise dependent. Matching is currently debated in studies involving stem cell transplantation, for instance, where randomising treatment is challenging or even infeasible. One approach to accounting for the correlation between matched partners is the stratified Cox model, however, one major drawback is reduced effective sample size. An alternative is the marginal Cox model, which estimates parameters in the same way as the classical Cox model. While this approach and the consistency of its estimator are discussed in the literature, a precise proof of consistency does not appear to be available. Once consistency is established, inference may be performed by resampling the matched pairs. This thesis aims to provide a proof of consistency. In addition to random right-censoring, the proof is extended to cover random left-truncation, censoring due to a competing risk and a combination of these mechanisms. Motivations for the latter come from registries on stem cell transplantation (competing risks) and from studying health policy interventions in calendar time (left-truncation). To assess the performance of the estimator in finite samples, a simulation study is conducted. Another simulation study compares the estimation of the variance of the parameter estimator using two methods: bootstrapping the matched pairs and applying a robust variance estimator that accounts for dependence.

3: Using Restricted Mean Survival Time under Proportional Hazards in a Non-Inferiority Randomised Trial with Time-to-Event Outcome

Matteo Quartagno, Matt Nankivell, Tim Morris, Ian White

MRC CTU at UCL, United Kingdom

Background Difference in Restricted Mean Survival Time (DRMST) has emerged as a promising summary measure in non-inferiority trials with time-to-event outcomes, offering potential advantages over the widely used Hazard Ratio (HR). However, there are remaining practical methodological questions to be answered before it can be used in trials that were originally designed using HR. These include the choice of horizon time (τ), and the conversion of non-inferiority margins. The PATCH trial serves as a case study to explore whether DRMST can enhance power in non-inferiority tests under proportional hazards scenarios and how these issues should be addressed.

Methods Simulations were conducted within the ADEMP framework, using PATCH trial design parameters and data generation mechanisms informed by observed data in the non-metastatic patients (M0) cohort. DRMST with various τ values was compared to HR in terms of power and type I error rates. Flexible parametric survival models estimated both DRMST and HR, and non-inferiority margins were converted under different assumptions about baseline hazard functions.

Results DRMST consistently demonstrated higher power than HR when non-inferiority margins were correctly matched to baseline hazard distributions. However, mismatched margins led to reduced power or inflated type I error rates. Converting the margin based on the observed survival distribution in the control arm seemed an acceptable compromise, not inflating type I errors substantially. Shorter τ values yielded greater power, and DRMST allowed analyses to be conducted with fewer events, enabling earlier trial conclusions.

Conclusions DRMST offers a robust alternative to HR in non-inferiority trials, with specific advantages in scenarios with proportional hazards. Careful selection of τ and margin conversion strategies is crucial for maximising power and maintaining statistical validity. Future research should address DRMST's performance in non-proportional hazards settings and refine guidance on its implementation in clinical trial design.

4: Variable Selection Methodology for Illness-Death Model with Interval Censored Data

Ariane Bercu¹, Agathe Guilloux^{2,3}, Cécile PROUST-LIMA¹, Hélène JACQMIN-GADDA¹

¹Inserm Research Center « Bordeaux Population Health », Bordeaux School of Public Health, CIC 1401-EC, Bordeaux University, 146 Rue Léo Saignat, 33000 Bordeaux cedex, France

²Inria Paris, F-75015 Paris, France

³Centre de Recherche des Cordeliers, INSERM, Université de Paris, Sorbonne Université, F-75006 Paris, France

Background Dementia is a chronic disease characterised by neurodegenerative processes and vascular brain injury. In prospective population-based cohorts, the diagnosis of dementia is usually interval-censored as it is made by a neuropsychologist at follow-up visits so that the exact time of dementia is unknown (1). The actual dementia status at death is also not always known as death may occur in between visits before a diagnosis of dementia can be made. The illness-death model for interval-censored data accounts for the uncertainty on the time of dementia and the probability of having dementia between the last visit without dementia and death (2). In this work, we proposed a new regularised estimation procedure for a high dimensional illness-death model with interval censored data, that performs variable selection.

Methods We considered a proximal gradient hybrid algorithm maximising the regularised likelihood with elastic-net penalty on the 3 transitions (healthy to dementia, health to death and dementia to death). Our algorithm simultaneously estimates all three transitions' regression parameters while having different penalty parameters on each transition. The performances of our algorithm were evaluated in simulations and compared to the alternative strategy of the standard competing risk approach that neglects interval censoring. The method was applied to the data of a French population-based cohort to identify the most important predictors of dementia risk in the elderly.

Results The illness-death model accounting for interval-censored data showed a very good ability to select relevant variables, in various scenarios. In comparison, the regularized competing risk model neglecting interval censoring tended to select irrelevant variables for the transition to dementia when associated with death. The proposed model also provided Mean Square Error of the probability of dementia very close to the oracle model that included only the relevant variables, and much smaller than the regularized model neglecting interval censoring. In the application, we identified regional brain volumes in addition to cognitive and socio-demographic markers as predictors of dementia risk.

Conclusion Our illness-death model extended to elastic-net penalty (available in the penidm R package) offers a promising solution to handle interval censored data.

1. Leffondré et al. Interval-censored time-to-event and competing risk with death: is the illness-death model more accurate than the Cox model? *International Journal of Epidemiology* 2013; 42:1177–86
2. Joly et al. A penalized likelihood approach for an illness-death model with interval-censored data: application to age-specific incidence of dementia. *Biostatistics* 2002; 3:433–43

5: Modelling the Time-Varying Effect of Hormonal Treatment on Metastasis-Free Survival among ER+ Breast Cancer Patients - a Natural History Modelling Approach

Letizia Orsini¹, Alessandro Gasparini^{1,2}, Kamila Czene¹, Keith Humphreys¹

¹Karolinska Institutet, Sweden

²Red Door Analytics AB, Stockholm, Sweden

Background Breast cancer treatment depends on tumour subtypes. In particular, patients with oestrogen receptor-positive (ER+) tumours are treated with hormonal therapy (either tamoxifen or aromatase inhibitors). In Sweden, the standard recommended treatment duration has historically been five years. However, current guidelines now suggest offering an additional five years of endocrine therapy to women at high risk of recurrence. This recommendation is supported by studies indicating that prolonged endocrine therapy may be associated with improved disease-free survival [1]. However, the impact of extended therapy on metastatic progression has not been quantified in a detailed way at the population level. In this article, we use a modelling approach to estimate the time-varying effect of hormonal treatment on the time to metastasis diagnosis. We then use it to compare 5-year and 10-year treatments for different tumour sizes.

Methods We incorporated the effect of endocrine therapy in a biologically inspired natural history model of breast cancer [2]. Individual tumour growth was modelled as exponential, with an inverse growth rate following a gamma distribution. We model the metastatic seeding with a non-homogenous Poisson process dependent on tumour size. We incorporate the treatment effect as a multiplicative factor of the inverse growth rate, allowing us to quantify its impact on tumour dynamics.

Results We fitted our model using a likelihood-based approach to a cohort of incident cases of 9716 patients diagnosed with invasive oestrogen receptor-positive breast cancer (ER+) between 2005 and 2020 who never received chemotherapy. 299 metastatic events occurred, with a median time to metastasis of 3.91 years [IQR: 2.36-6.15]. Based on our model estimates, for patients with 15mm and 20mm tumours the gains in 10-year metastasis-free survival, from receiving ten years instead of five years of hormonal treatment are expected to be approximately 1.5 and 3 percentage points, respectively.

Conclusion Our natural history model quantifies the impact of prolonged hormonal treatment on metastatic events in ER+ breast cancer patients. The results demonstrate a significant

Abstracts of Contributed Talks

reduction in tumour growth rates during treatment, supporting the extension of endocrine therapy to 10 years for patients with large tumours.

[1] Zeng, E., et al. (2022). Determinants and effectiveness of extending the duration of adjuvant hormone therapy beyond 5 years in patients with breast cancer. *Cancer Research*, 82(19), 3614-3621.

[2] Gasparini, A., & Humphreys, K. (2022). Estimating latent, dynamic processes of breast cancer tumour growth and distant metastatic spread from mammography screening data. *Statistical methods in medical research*, 31(5), 862-881.

Causal Inference in Time-Varying Settings

Wednesday, 2025-08-27 16:00 - 17:30, Biozentrum U1.101

Chair: Oliver Dukes

1: The Causal Effect of Gold Standard Midwifery Staffing on the Occurrence of Spontaneous Vaginal Births – a Target Trial Emulation

Luisa Eggenschwiler^{1,2}, Valerie Smith³, Michael Simon^{1,2}, Giusi Moffa⁴

¹Institute of Nursing Science, University of Basel, Switzerland

²Chief Medical and Nursing Office, University Hospital Basel, Switzerland

³School of Nursing, Midwifery, and Health Systems, University College Dublin, Ireland

⁴Department of Mathematics and Computer Science, University of Basel, Switzerland

Background Studies have suggested that optimal midwifery staffing is associated with spontaneous vaginal births. The gold standard is a midwife-to-woman ratio of one-to-one during established labour and birth. In a clinical site, where the gold standard is established, it is not ethical to conduct a randomised controlled trial with less than standard care. In the real world though, it is not always possible to adhere to the gold standard due to high variation in care demand and ineffective measures to address high demand. Observational data will thus consist of gold standard and non-gold standard cases. To estimate causal effects, a target hypothetical pragmatic randomised trial was conceptualised with the routinely collected hospital data. The aim was to determine the causal effect of gold standard midwifery staffing compared to less than gold standard midwifery staffing on the occurrence of spontaneous vaginal births.

Methods The target trial was emulated with routine hospital data from one tertiary hospital from 01 January 2019 to 31 December 2022. The exposure was defined as the proportion of time with one-to-one care during active labour, with 100% midwifery staffing as gold standard. All women in active labour were eligible to be included. Women with a planned caesarean section, breech birth, multiple pregnancy and stillbirth were excluded. Women assigned to each staffing exposure were assumed to be comparable conditional on baseline covariates. We considered hypertensive disorders, diabetes, gestational age, parity, maternal age, country of birth, birth weight and labour induction as confounding variables and adjusted for them with inverse probability weighting. We used marginal structural models to calculate the total effect based on the per-protocol policy. Ethical approval to access the routine hospital data has been granted by the local ethics committee.

Results In total 6,602 cases were included in the analysis and 16.2% ($n=1,072$) were exposed to gold standard midwifery staffing. Overall, 61% ($n=4031$) had a spontaneous vaginal birth. Women receiving gold standard midwifery staffing are 5.7% (CI 2.3% – 8.9) more likely to have a spontaneous vaginal birth than women not receiving gold standard midwifery staffing.

Conclusion The target trial emulation confirmed what cross-sectional studies already have indicated. Women receiving gold standard midwifery staffing are more likely to have a spontaneous vaginal birth. The target trial emulation can be used as a blueprint for further research of causal links between nurse and midwifery staffing and outcomes.

2: A New Encoding of Time-Varying Treatment Regimes for the Study of Sequential per-Protocol Effects

Ignacio Gonzalez-Perez, Mats Julius Stensrud

EPFL, Switzerland

We describe an alternative encoding of time-varying treatment strategies. The motivation for this encoding is to simplify the exposition of identifiability assumptions in many settings of practical interest. As a clinically relevant running example, we consider sequential per-protocol effects, including sequential separable effects. The study of these effects would complement the traditional intention-to-treat analyses of an RCT. We derive new identification results for these per-protocol parameters, which can be described as conditional independencies in causal Directed Acyclic Graphs (DAGs) and Single World Intervention Graphs (SWIGs). Furthermore, we propose several estimators, including one that is semi-parametrically efficient and double-robust. These results are illustrated through an analysis of the Systolic Blood Pressure Intervention Trial (SPRINT), where we estimate a new type of time-varying separable effects, with a clear clinical interpretation as the per-protocol separable effect of taking a modified blood pressure treatment on acute kidney injury.

3: Optimal Sequential Decision-Making with Initiation Regimes

Julien David Laurendeau¹, Aaron Leor Sarvet², Mats Julius Stensrud¹

¹Swiss Federal Institute of Technology Lausanne (EPFL)

²University of Massachusetts Amherst

Consider an optimal dynamic treatment regime, correctly identified from a perfectly executed sequentially randomised experiment. Even when the experimental results are generalisable to a future target population, there is no guarantee that the optimal regime outperforms human decision-makers; human experts can do better than the optimal regime whenever they have access to relevant information beyond the covariates recorded in the experiment. Motivated by this fact, we derive results on a new class of regimes called initiation regimes. These regimes follow human decision-makers until it is more beneficial to initiate a sequential optimal regime from that point onward, and are guaranteed to outperform both entirely human and entirely algorithmic decision-makers, e.g., based on reinforcement learning algorithms. Furthermore, we present modified experimental designs that identify the best initiation regimes, show how the best initiation regime can be identified from classical observational data with commonly invoked assumptions, and give estimation and statistical inference methodology for these regimes. To illustrate the practical utility of our methods, we consider initiation regimes in a case study on back pain treatment.

4: Evaluating the Effect of Lung Transplantation: a Case Study in Sequential Emulated Trials with Time Varying Confounding

Iqraa Meah, François Petit, Raphaël Porcher

CRESS, Methods team

Lung transplantation was a critical intervention for extending the lifespan of individuals with cystic fibrosis. Since transplant assignment cannot be randomized, evaluating treatment effectiveness relies on observational data. Such data—such as those provided by the United Network for Organ Sharing (UNOS)—offer a valuable opportunity to emulate a target trial. This methodology is widely used to investigate causal relationships using observational data, which inherently contain biases. One major source of bias is confounding due to non-random treatment assignment. Additionally, improper implementation of the emulated trial framework can introduce further biases, such as immortal time bias, which complicates the estimation of treatment effects. Correcting for this bias, however, can induce informative censoring bias, adding another layer of complexity to the analysis.

In this work, we use UNOS data as a case study to develop a methodological framework for emulating target trials in the context of lung transplantation. We address the challenges associated with different types of bias, leading to a sequence of target trials that incorporate

time-dependent matching based on the lung allocation score (LAS), the primary known confounder affecting treatment assignment. Specifically, our design involves setting a sequence of time landmarks (e.g., weekly follow-ups) to minimize the time between follow-up initiation and transplantation for the treated group. Each treated individual is matched with a control - who is on the waiting list at the current time - ensuring similarity based on LAS. If a control later receives a transplant, they are artificially censored and subsequently included in the treated population in a later trial. To adjust for this informative censoring, we apply the Inverse Probability of Censoring Weighting (IPCW) method with time-varying weights. We present our results as survival curves aggregated from Kaplan-Meier estimators fitted on each weekly trial.

Finally, we explore a theoretical question regarding the convergence of such an aggregated estimator when based on dependent data, as individuals on the waiting list can participate in multiple trials. We propose to investigate the level of dependency as a function of the total number of individuals available for the study through simulations, analyzing the effects of sequentialization, matching, and censoring.

5: Unraveling Time-Varying Causal Effects of Multiple Exposures: a Novel Approach Integrating Functional Data Analysis into the Multivariable Mendelian Randomization Framework

Nicole Fontana^{1,2}, Piercesare Secchi¹, Emanuele Di Angelantonio², Francesca Ieva^{1,2}

¹MOX, Department of Mathematics, Politecnico di Milano, Milan, Italy

²Health Data Science Research Centre, Human Technopole, Milan, Italy

Background The causal effects of exposures often vary across an individual's lifetime, with certain periods exerting a greater influence on health outcomes or revealing the long-term consequences of risk factors. Capturing these time-varying causal effects provides valuable insights into underlying mechanisms and supports the development of effective health interventions. Multivariable Mendelian Randomization (MVMR) offers a robust framework to estimate the direct effect of any exposure on an outcome, accounting for the influence of other concurrent exposures. However, methods for accounting for time-varying exposures remain limited in the literature.

Methods We propose a Multivariable Functional Mendelian Randomization (MVFMR) approach to estimate the direct effects of multiple time-varying exposures on an outcome. Our method builds on the framework introduced by [1]. First, we apply functional principal component analysis (FPCA) to reduce dimensionality by extracting low-dimensional factors

from exposure trajectories. Second, we introduce a data-driven feature selection step to determine the optimal number of functional principal components (FPCs), eliminating the need for predefined selection criteria. Finally, we develop the MVFMR model to assess the direct causal associations of multiple time-varying exposures with the outcome, providing a comprehensive evaluation of their effects across an individual's lifetime.

Results Through simulations, we demonstrate that our feature selection approach effectively identifies the most accurate functional form of the association between time-varying exposures and outcomes. The strength of this approach is that it selects the optimal number of principal components without relying on *a priori* definitions based on the variability explained by them, criteria that our simulations demonstrate to be unreliable. Furthermore, we demonstrate that our method outperforms separate models in accurately estimating the time-varying effects of multiple exposures when all are associated with the outcome. We applied the proposed methodology to investigate the impact of time-varying genetically predicted systolic blood pressure and LDL cholesterol on the risk of coronary artery disease, using data from the UK Biobank.

Conclusions This study highlights the importance of integrating functional data analysis within the Mendelian Randomization framework to understand how risk factors evolve over a lifetime and estimate their causal effects. This approach enables the identification of critical exposure periods, providing valuable clinical insights that can inform targeted healthcare strategies.

- [1] Tian, H., Patel, A., & Burgess, S. (2024). Estimating Time-Varying Exposure Effects Through Continuous-Time Modelling in Mendelian Randomization. *Statistics in medicine*, 43(27), 5166–5181. <https://doi.org/10.1002/sim.10222>

Prediction / Prognostic Modelling 4

Wednesday, 2025-08-27 16:00 - 17:30, ETH E27

Chair: Richard D Riley

1: Adapting Existing Sample Size Calculations for Developing Risk Prediction Models to Control for Model Stability

Gareth Ambler, Menelaos Pavlou, Rumana Z Omar

University College London, United Kingdom

Background The use of recently proposed sample size calculations can lead to more reliable risk prediction models. These calculations can determine the required sample size to ensure that the expected calibration slope (CS), a measure of model overfitting, will meet some target value (often 0.9) when models are fitted using MLE; a perfectly calibrated model has a CS of 1. These calculations require information on the number of predictors, outcome prevalence and model strength.

Methods In practice, the observed CS will vary around the target value. This aspect of model performance, model stability, is not accounted for in existing calculations. We use simulation to investigate model stability when varying the number of predictors (p), outcome prevalence and model strength (c-statistic). We quantify model stability using the Probability of Acceptable Calibration (PAC), defined here as achieving an observed CS within [0.85-1.15]. We also investigate the performance of a simple (post-estimation) uniform shrinkage approach (using bootstrapping) which may be useful in some scenarios. Finally, we propose an adaptation of existing sample size calculations to control model stability and ensure that PAC is sufficiently high; here we aim for PAC=75%.

Results When adhering to existing sample size recommendations, the variability in the observed CS increased substantially with decreasing p (although CS=0.9 was achieved on average). Consequently, PAC was often low, particularly for $p < 10$. Applying simple uniform shrinkage led to much higher PAC unless there were very few predictors ($p \leq 5$). Our proposed adaptation resulted in higher sample sizes than those currently recommended for $p < 15$, and similar sizes for higher p .

Conclusions Sample size calculations for the development of prediction models should take account of model stability. Applying post-estimation shrinkage at the existing recommended sizes may be beneficial unless the number of predictors is very small.

References Riley et al. 2020. Calculating the sample size required for developing a clinical prediction model. *British Medical Journal*, <https://doi.org/10.1136/bmj.m441>.

Pavlou et al. 2024. An evaluation of sample size requirements for developing risk prediction models with binary outcomes. *BMC Medical Research Methodology*, <https://doi.org/10.1186/s12874-024-02268-5>.

Riley & Collins 2023. Stability of clinical prediction models developed using statistical or machine learning methods. *Biometrical Journal*, <https://doi.org/10.1002/bimj.202200302>.

2: Optimal Methods for Handling Continuous Predictors in Clinical Prediction Model Development: Balancing Flexibility and Stability

Phichayut Phinyo¹, Pakpoom Wongyikul¹, Noraworn Jirattikanwong¹, Natthanaphop Isaradech², Wachiranun Sirikul², Wuttipat Kiratipaisarl², Noppadon Seesawan³, Suppachai Lawanaskol⁴

¹Department of Biomedical Informatics and Clinical Epidemiology (BioCE), Faculty of Medicine, Chiang Mai University, Chiang Mai, Thailand

²Department of Community Medicine, Faculty of Medicine, Chiang Mai University, Chiang Mai, Thailand

³Department of Emergency medicine, Lampang Hospital, Muang District, Lampang, Thailand

⁴Chaiprakarn Hospital, Chaiprakarn, Chiang Mai, Thailand

Background Prediction stability is increasingly recognised as essential for ensuring reliable and reproducible model development. While dataset size and algorithm choices affect stability, the impact of specific modelling decisions, particularly in handling continuous predictors, is less well understood. This study examines how different methods for handling continuous predictors influence stability.

Methods A dataset of 19,418 patients, previously used to develop a prediction model for hospital admission, was randomly sampled to create five datasets of different sizes [0.2, 0.5, 1, 2, and 5 times the base size]. The base size represented the minimum sufficient sample size for developing models. We defined six continuous candidate predictors, with most showing non-linear association with the endpoint. Six approaches for handling continuous predictors were compared: (1) DICO – dichotomisation, (2) CAT – categorisation into tertiles, (3) LINEAR – assuming a linear relationship, (4) QUAD – assuming a quadratic relationship, (5) MFP – multivariable fractional polynomial transformations, and (6) XGBoost

– extreme gradient boosting. Logistic regression was used for methods (1) to (5). Prediction stability was assessed using the bootstrap procedure proposed by Riley and Collins. Optimism-corrected area under the curves (AUCs) and calibration slopes were estimated to assess model performance. A modelling approach was considered highly stable if 90% of predictions had a mean absolute prediction error (MAPE) of 5%.

Results At the base size, DICOH, LINEAR, and QUAD produced highly stable predictions with similar levels of calibration. However, DICOH exhibited lower AUCs than the other two. While MFP and XGBoost demonstrated higher AUCs than the others, their predictions lacked stability, and their calibration was poor. At larger sample sizes ($2 \times \text{Base}$ and $5 \times \text{Base}$), all methods achieved high stability. LINEAR, QUAD, and MFP outperformed DICOH and CAT in AUCs. Although XGBoost produced stable predictions with high AUCs, but significant miscalibration persisted. With a smaller-than-sufficient sample sizes ($0.2 \times \text{Base}$ and $0.5 \times \text{Base}$), LINEAR and DICOH demonstrated better stability than more complex methods (i.e. QUAD, MFP and XGBoost).

Conclusions Methods for modelling continuous predictors should be chosen based on sample size and the trade-off between discrimination, calibration, and stability. For sufficiently large sample sizes, LINEAR and QUAD are preferred. MFP can yield higher AUCs but requires a substantially larger sample size for stability. XGBoost may not be an optimal choice even with a large sample size if calibration is a priority. For small datasets, LINEAR—and even DICOH—may be preferable to more complex, flexible methods to ensure maximal stability.

3: Do Stable Performance Metrics Guarantee Stable Model Predictions?

Natthanaphop Isaradech¹, Phichayut Phinyo², Wuttipat Kiratipaisarl¹, Pakpoom Wongyikul², Noraworn Jirattikanwong², Wachiranun Sirikul¹

¹Department of Community Medicine, Faculty of Medicine, Chiang Mai University, Chiang Mai, Thailand

²Department of Biomedical Informatics and Clinical Epidemiology (BioCE), Faculty of Medicine, Chiang Mai University, Chiang Mai, Thailand

Introduction Clinical prediction models are generally developed using either statistical models or machine learning algorithms to support clinical decision-making in critical situations. These models are typically evaluated based on their predictive performance, particularly in terms of discrimination and calibration. Concerns about the stability of these performance metrics have led to the need for internal validation and the estimation of model optimism. More recently, attention has shifted towards the stability of model predictions, which is more

relevant to clinical practice, as unstable predictions can lead to inconsistent risk estimations and ultimately impact patient outcomes. This study aims to determine the association between performance metrics optimism and stability metrics (e.g., the classification instability index (CII)).

Methods We replicated the development of a previously published prediction model for mortality using the GUSTO-I dataset. Ten scenarios of sample sizes (500, 1000, 2000, 3000, 4000, 5000, 6000, 10000, 20000, and 40830) were drawn through stratified random sampling. Seven candidate predictors were defined, and logistic regression was used for model derivation. Each dataset underwent the same development process and was evaluated for predictive performance and stability using 200 bootstrap resamples. We estimated and compared mean absolute prediction error (MAPE) and the classification instability index (CII) with the estimated optimism of performance metrics, such as the area under the ROC curve (AuROC).

Results We found that models trained on different sample sizes exhibited similar discrimination performance, with a good AuROC (mean 0.73, 95% CI: 0.70–0.76) and favorable optimism (mean 0.017, 95% CI: 0.009–0.025). However, mean AUC optimism showed a clear downward trend and decreasing standard deviations as the sample size increased. This pattern was also observed in stability metrics such as CII, MAPE, estimated risk (relative to the original model), and calibration, despite the models generally having low optimism. The results also showed MAPE instability and values exceeding 5%, with unstable calibration, especially at high predicted probabilities in sample sizes of 1,000–5,000. MAPE began to stabilize and remain low at 10,000 samples.

Conclusion Stable discriminative performance metrics and low optimism from internal validation do not necessarily indicate stability in model prediction and calibration. Model developers should be cautious about relying solely on consistent discrimination performance and low optimism, particularly when the model is trained on a small sample size.

4: The Imbalance Dilemma: Can Class Imbalance Corrections Stabilize Clinical Prediction Models?

Wachiranun Sirikul^{1,2}, Natthanaphop Isaradech¹, Wuttipat Kiratipaisarl¹, Phichayut Phinyo², Pakoom Wongyekul², Noraworn Jirattikanwong²

¹Department of Community Medicine, Faculty of Medicine, Chiang Mai, Thailand

²Department of Biomedical Informatics and Clinical Epidemiology (BioCE), Faculty of Medicine, Chiang Mai, Thailand

Background Class imbalance is a common problem in developing clinical prediction models (CPMs), often leading to a prediction paradox. A variety of methods for correcting data imbalance have been proposed and implemented to improve the development of CPMs. However, correcting for imbalance could potentially degrade model performance and exacerbate bias by increasing overfitting. In this study, we investigated how imbalance correction influenced the performance of logistic regression models, focusing on prediction stability using the Gusto dataset.

Methods Model development and internal validation were done using different methods to correct for class imbalance (none, SMOTENC, BorderlineSMOTE, and ADASYN). This was done with the imbalanced-learn library in Python (version 3.12) and with sample sizes of 500, 1000, 2000, 3000, and 40830. The smallest sample scenario was determined based on the minimum sample size required for developing a multivariable classification model by Riley RD et.al. CPMs were developed using penalised logistic regression with hyperparameter tuning using grid search with 10-fold cross-validation via sklearn. Model performance and prediction stability were evaluated using 200 bootstrap samplings to obtain the area under received operating characteristic curves (AuROCs) with optimism corrected, calibrations, mean absolute prediction error (MAPE), and classification instability indices (CII).

Results In the full sample scenario, the AuROCs with optimism correction for the models without imbalance correction, SMOTENC, BorderlineSMOTE, and ADASYN were 0.768, 0.802, 0.803, and 0.789, respectively. All models using imbalance corrections showed better prediction stability in discrimination, calibration, MAPE, and CII compared to the model without an imbalance correction. In the minimal required sample scenario ($n=500$), the models with imbalance corrections, except for ADASYN, improved model discrimination compared to the model using the original data. Despite the limited sample size affecting all models' calibration and prediction stability, the original data model had the most stable predictions. With more saturated data scenarios ($n=3000$), the findings of performance and stability were consistent with the full sample scenario.

Conclusion Our study demonstrated that using oversampling techniques for imbalance correction in simple clinical data and standard statistical models not only improved model discrimination but also enhanced the calibration and stability of CPMs in large sample sizes. However, applying imbalance corrections in small samples, such as the minimal required sample, should be approached with caution. The effects of imbalance corrections can be attributed to the trade-off between introducing optimal bias to make the model more suitable for the minority class and applying optimal model penalisation to mitigate overfitting.

5: Prediction with Logistic Regression in Binary Class Imbalance: Comparing Re-Sampling Techniques with Threshold Probability Assignment under Varying Predictive Covariates

Henk van der Pol^{1,2}, Ragnhild Sørum Falk³, Marta Fiocco^{2,4,5}, Arnoldo Frigessi⁶, Euloge Clovis Kenne Pagui^{3,6}

¹Department of Medical Oncology, Leiden University Medical Center, the Netherlands

²Mathematical Institute, Leiden University, the Netherlands

³Oslo Centre for Biostatistics and Epidemiology, Oslo University Hospital, Oslo, Norway

⁴Princess Máxima Centre for Paediatric Oncology, Utrecht, the Netherlands

⁵Department of Biomedical Data Science, Section Medical Statistics, Leiden University Medical Centre, Leiden, the Netherlands

⁶Oslo Centre for Biostatistics and Epidemiology, University of Oslo, Oslo, Norway

Background Prediction of rare events in binary classification is a well-studied topic in biostatistics, which has gained renewed interest in the context of Machine Learning (ML). Class imbalance is often addressed in the ML-community through re-sampling techniques such as random over sampling, random under sampling, and synthetic minority oversampling technique (SMOTE). However, recent research has shown that re-sampling techniques are not always necessary and may even be harmful to clinical prediction [1]. Nevertheless, little interest is shown how a proper threshold probability assignment strategy may overcome this issue. Importantly, this study explores how the strength of the covariates and the correct specification of model affect the prediction performance in an imbalanced setting. The aim of this research is to investigate the impact of class imbalance correction strategies on the performance of logistic classifiers, when covariates have varying degree of predictive power.

Methods We conducted a Monte Carlo simulation, based on a logistic regression model with various settings such as, prevalence of disease, re-sampling techniques, model specification and strength of covariates. The predictive performance is mainly assessed by the precision and recall (sensitivity) measures. We compare the re-sampling techniques with several threshold probabilities.

Results In all simulation scenarios, proper threshold probability assignment strategy have comparable or better prediction performance compared to re-sampling techniques. Specifically, the threshold that maximizes the area under the ROC curve (AUC) returns same level of precision compared to the SMOTE re-sampling technique. However, the recall is on average, 10% greater. Moreover, these differences are enlarged with the increase strength of a predictive variable. Lastly, we reinforce current findings in which we show that un-correcting the dataset returns a higher AUC and Area under the precision and recall curve in imbalanced data, compared to re-sampling techniques.

Conclusion Proper threshold probability assignment strategy outperforms re-sampling techniques in imbalanced setting. Focus should continue on optimal threshold probabilities based on the outcome of interest, investigate predictive variables.

Reference [1] van den Goorbergh, R., van Smeden, M., Timmerman, D., & Van Calster, B. (2022). The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. *Journal of the American Medical Informatics Association*, 29(9), 1522-1531.

Meta-Analysis 3

Wednesday, 2025-08-27 16:00 - 17:30, ETH E23

Chair: Ulrike Held

1: Meta-Analysis of Time-to-Event Outcomes with Non-Proportional Hazards - A Time-Varying Hazard Ratio Approach

Keith R Abrams¹, Rhiannon K Owen²

¹University of Warwick, United Kingdom

²Swansea University Medical School, United Kingdom

Background

Often when undertaking meta-analyses of time-to-event (TTE) outcomes, especially in Health Technology Assessment (HTA), a hazard ratio scale is used. However, issues arise when there is evidence of non-proportional hazards in some of the trials/studies included. When Individual Patient Data (IPD) are available or have been re-created from published Kaplan-Meier curves a number of methods have been advocated, including; flexible parametric models, piecewise exponential models, fractional polynomial models, and Restricted Mean Survival Time (RMST) models. However, their use has been limited by either their complexity and/or the ease with which their results can be incorporated into an economic decision model in order to assess cost-effectiveness.

An alternative approach is to assume a treatment-log(time) interaction within a Cox proportional hazards model in each trial/study, thus allowing the log HR to vary linearly with respect to log(time), and to then undertake a bivariate meta-analysis of the resulting treatment and interaction coefficients, so that an overall time-varying HR can be obtained.

Methods

A treatment-log(time) approach was applied to an IPD meta-analysis of 20 trials, involving 4,069 patients, of chemotherapy compared to Standard of Care (SoC) for advanced recurrent gastric cancer undertaken by the Global Advanced/Adjuvant Stomach Tumor Research International Collaboration (GASTRIC) Group, and in which Progression-free Survival (PFS) was an outcome with follow-up up to 6.8 years (2500 days). This approach was compared with a standard random effects meta-analysis of the trial-specific log HRs.

Results

Of the 20 trials in the meta-analysis 5 displayed evidence of non-proportional hazards for PFS. A standard random effects meta-analysis of HRs (undertaken on a log scale) yielded a pooled HR of 0.78 (95% CI: 0.71 to 0.86). Undertaking a bivariate random effects meta-analysis of the treatment and treatment-log(time) trial-specific coefficients produced a pooled interaction effect of +0.10 (95% +0.002 to +0.19) $P=0.04$ on a log hazard scale. The resulting HRs estimated at 250, 500, 750 and 1000 days were 0.85 (95% CI: 0.78 to 0.93), 0.91 (95% CI: 0.80 to 1.04), 0.95 (95% CI: 0.81 to 1.11) and 0.97 (95% CI: 0.81 to 1.17) respectively.

Conclusion

A treatment-log(time) interaction approach to the meta-analysis of TTE outcomes when the proportional hazards assumption appears not to hold for at least some of the studies included produces a simple and intuitive solution which can be readily incorporated into an economic decision model. Further extension to both a network meta-analysis setting and a fully Bayesian one-stage model is also possible.

2: Flexing the Curve: Comparing Spline and Fractional Polynomial Models in Network Meta-Analysis of Survival Outcomes

Suraj Balakrishna, Justin Chumbley, Natalia Popova, Shahrul Mt-Isa

MSD Innovation & Development GmbH, Switzerland

Background There is a growing demand to perform network meta-analysis (NMA) in health technology assessment (HTA) submissions, especially to meet the new EU-HTA requirements. Time-to-event clinical endpoints are commonly used in many disease areas. In NMAs with multiple studies, there is a high chance that proportional hazards (PH) assumption is not met for at least one study. In such situations, flexible methods for survival modelling are more suitable. Fractional polynomials (FPs) and Splines are commonly used flexible parametric models when PH assumption fails. FPs have a simpler mathematical form and require fewer parameters but may not capture all possible non-linear relationships. In contrast, splines are more flexible and can fit to more complex patterns, although they risk overfitting due to their high flexibility and consequently may not be suitable for extrapolation. However, recent developments in spline models have addressed the risk of overfitting. In this study, we assess the performance of FP and spline models.

Methods We simulate time-to-event data for multiple studies within a NMA network following different underlying distributions including (a) simple parametric distributions (b) complex distributions to account for delayed response to treatment and the existence of long-term survivors. We fit various FP and spline models to this simulated network to compare their predictive and extrapolation performance for new data simulated from the same underlying data generating process.

Outlook: This study provides insights into the relative performance of FP and spline models in NMA and may help one to choose the appropriate flexible parametric modelling approach for time-to-event data.

3: Fixed and Random Effect Meta-Analysis for Competing Risks: a How-to Guide for Aggregated and Individual Participant Data

Matthias Klimek¹, Matthias Schmid², Anja Rüten-Budde³, Andreas Ziegler^{1,4,5,6}

¹Cardio-CARE, Medizincampus Davos, Davos, Switzerland

²Institute of Medical Biometry, Informatics and Epidemiology, Faculty of Medicine, University of Bonn, Bonn, Germany

³Anja Juana Rüten-Budde, Statistician Next Door, Leoben, Austria

⁴Department of Cardiology, Hochgebirgsklinik Davos, Davos, Switzerland

⁵Center for Population Health Innovation (POINT), University Heart and Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

⁶School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Pietermaritzburg, South Africa

Meta-analysis summarises the results of multiple independent clinical trials. For survival outcomes, both fixed effect and random effect meta-analyses are traditionally carried out using aggregated data obtained from published effect estimates. However, relevant data may not have been published for specific analyses, such as subgroups of interest, or administrative censoring may have occurred at different time points. In such cases, individual participant data (IPD) meta-analyses may be performed. The aim of this work is to provide a practical guide for the conduct of IPD fixed effect and random effect meta-analyses for survival outcomes in the presence of competing events. For the random effect analysis, we model optimal correlated frailties from the gamma distribution representing unobserved covariates at the cluster level, thus allowing for correlations between the different competing events within the studies. This presentation builds on previous work by Meddis et al. (2020 Biom J) and Rueten-Budde et al. (2019 Stat Med). It provides competing risks estimators for both cause-specific and subdistribution hazard ratios. For illustration, the IPD data from

Meddis et al. are re-analysed. Specifically, data from 23 randomised controlled trials with a total of 4552 patients suffering from nasopharyngeal carcinoma and two competing events are considered. In summary, fixed effect and random effect meta-analyses can be performed with ease on both aggregated and IPD data.

4: The Transmission Blocking Activity of Artemisinin-Combination, Non-Artemisinin, and 8-Aminoquinoline Antimalarial Therapies: a Network Meta-Analysis.

Jordache Ramjith¹, Leen N. Vanheer², Almahamoudou Mahamar³, Merel J. Smit¹, Kjerstin Lanke¹, Michelle E. Roh⁴, Koualy Sanogo³, Youssouf Sinaba³, Sidi M. Niambélé³, Makonon Diallo³, Seydina O. Maguirega³, Sekouba Keita³, Siaka Samake³, Ahamadou Youssouf³, Halimatou Diawara³, Sekou F. Traore³, Roly Gosling^{5,6}, Joelle M. Brown⁶, Chris Drakeley², Alassane Dicko³, Will Stone², Teun Bousema¹

¹Department of Medical Microbiology and Radboud Center for Infectious Diseases, Radboud University Medical Center, Nijmegen, the Netherlands

²Department of Infection Biology, London School of Hygiene & Tropical Medicine, London, UK, WC1E7HT

³Malaria Research and Training Centre, Faculty of Pharmacy and Faculty of Medicine and Dentistry, University of Sciences Techniques and Technologies of Bamako, Bamako, Mali

⁴Institute for Global Health Sciences, University of California, San Francisco, CA, USA.

⁵Department of Disease Control, London School of Hygiene and Tropical Medicine, London UK.

⁶Department of Epidemiology and Biostatistics, University of California, San Francisco, CA, USA.

Background Interrupting human-to-mosquito transmission is important for malaria elimination strategies as it can reduce infection burden in communities and slow the spread of drug resistance. Antimalarial medications differ in their efficacy in clearing the transmission stages of *Plasmodium falciparum* (gametocytes) and in preventing mosquito infection. Here we present a combined analysis of six trials conducted at the same study site with highly consistent methodologies that allows for a direct comparison of the gametocytocidal and transmission-blocking activities of fifteen different antimalarial regimens or dosing schedules.

Methods and findings: Between January 2013 and January 2023, six clinical trials with transmission endpoints were conducted at the Clinical Research Centre of the Malaria Research and Training Centre of the University of Bamako in Mali. These trials tested Artemisinin-

Combination Therapies (ACTs), non-ACT regimens and combinations with 8-aminoquinolines. Participants were males and non-pregnant females, between 5-50 years of age, who presented with *P. falciparum* mono-infection and gametocyte carriage by microscopy. Blood samples were taken before and after treatment for thick film microscopy, infectivity assessments by mosquito feeding assays and molecular quantification of gametocytes. A network meta-analysis (NMA) was performed to combine direct and indirect effects of treatment arms across studies. This analysis quantified changes in mosquito infection rates and gametocyte densities within treatment arms at 2 days, 7 days, and 14 days post-regimen relative to baseline (day 0). These quantified relative changes within arms were also compared between arms. In a pooled analysis of 422 participants, we observed substantial differences between antimalarials in gametocytocidal and transmission-blocking activities, with artemether-lumefantrine (AL) being significantly more potent at reducing mosquito infection rates within 48 hours than dihydroartemisinin-piperaquine (DHA-PPQ), artesunate-amodiaquine (AS-AQ), sulfadoxine-pyrimethamine plus amodiaquine (SP-AQ) and pyronaridine-artesunate (PY-AS) ($p<0.0001$). The addition of single low dose primaquine (SLD PQ) accelerated gametocyte clearance and led to a significantly greater reduction in mosquito infection rate within 48-hours of treatment for each ACT, while an SLD of the 8-aminoquinoline tafenoquine (TQ) showed a delayed but effective response compared to SLD primaquine.

Conclusions We found marked differences among ACTs and single low-dose 8-aminoquinoline drugs in their ability and speed to block transmission. The findings from this analysis can support treatment policy decisions for malaria elimination and be integrated into mathematical models to improve the accuracy of predictions regarding community transmission and the spread of drug resistance under varying treatment guidelines.

5: Meta-Analysis of Diagnostic Test Accuracy with Multiple Disease States: Combining Stage-Specific Accuracy Data with Descriptive Statistics

Efthymia Derezea¹, Gabriel Rogers², Nicky Welton¹, Hayley E Jones¹

¹Population Health Sciences, Bristol Medical School, University of Bristol, UK

²Manchester Centre for Health Economics, University of Manchester, UK

Introduction Standard meta-analysis of diagnostic test accuracy assumes a binary classification of participants, i.e. diseased or healthy, and estimates the (overall) sensitivity and specificity. Sometimes, however, we need estimates of accuracy for multiple disease states. For example, it is important to know the ability of a test to detect cancer at different stages (early, advanced etc.), since there may be greater potential benefit of diagnosing at an earlier stage, when more amenable to treatment. If sufficient studies report stage-specific sensitivity

(“subgroup data”), we might pool these using standard methods. However, often very few studies report this. In a systematic review of the accuracy of tests to detect hepatocellular carcinoma (HCC) among people with cirrhosis, we found that many more studies reported the proportion of detected HCCs that were at each stage (“baseline proportions”), however.

Methods We propose a method for obtaining meta-analysed accuracy estimates for multiple disease states, combining subgroup and baseline data. Where stage-specific data are not reported, we assume overall sensitivity is an average of stage-specific sensitivities, weighted by the baseline proportions. Study-level random effects allow for heterogeneity and potential between-study correlations among the disease states. We further extend this approach to continuous tests reporting accuracy at multiple thresholds, building on the model of *Jones et al, 2019*. This produces pooled estimates of accuracy for multiple disease states, at any diagnostic threshold, based on a combination of subgroup and baseline data.

Results By applying this method to simulated and real datasets from the HCC systematic review, we show that the model can produce results with increased precision compared to those obtained by meta-analysing stage-specific data alone. For one continuous test, alpha-fetoprotein (AFP), only four and five studies respectively reported sensitivity to detect HCCs at a ‘very early’ or ‘early’ stage – the most clinically relevant quantities – and none of these reported sensitivity across multiple thresholds, rendering results from subgroup data alone meaningless. Our model allowed the inclusion of 33 more studies and produced estimates of sensitivity to detect cancer of each stage, and specificity, across all thresholds.

Conclusions This approach makes the most of descriptive statistics reported in many test accuracy studies to supplement the more informative, but less often reported, subgroup data. This allows us to produce pooled estimates of accuracy for each disease stage or level of severity, across all thresholds.

Observational/Real-World Data 2

Wednesday, 2025-08-27 16:00 - 17:30, ETH E21

Chair: Sabine Hoffmann

1: Group Measurement Invariance Assessment with Item-Level Latent Variable Models: A Comparative Simulation Study of Two Methods

Myriam Blanchin¹, Odile Stahl¹, Yseulys Dubuy¹, Véronique Sébille^{1,2}

¹Nantes Université, Université de Tours, INSERM, UMR1246 SPHERE « methodS in Patient-centered outcomes and HEalth ResEarch », Nantes, France

²CHU Nantes, DRCI, Methodology and Biostatistics Department, Nantes, France

Introduction Measurement invariance assessment is essential when comparing health-related quality of life between groups as it can bias mean comparisons and obfuscate the interpretation of intervention effect. Ordinal item responses to quality of life questionnaires can be analyzed with partial-credit models. Among the various methods for invariance assessment, also known as Differential Item Functioning (DIF) analysis, only a few are available to analyze DIF in multiple groups of patients. The objective was to compare two methods for DIF analysis with latent regression partial-credit models across three groups of patients.

Methods Ordinal responses of three groups of patients were simulated with a partial-credit model. Sample sizes, number of items and response categories, DIF patterns (DIF in none, two pairs of groups or all groups), and group effect values all varied in the simulation scenarios.

The anchor selection method consists in i/ an iterative Wald test procedure to identify a stable set of anchor items (DIF-free items) starting from an unrestricted model (DIF on all items), and ii/ a DIF refinement of all DIF items at once to determine the groups affected by DIF and the type of DIF. The DIF items identification method performs first a likelihood-ratio test between the fully invariant and the unrestricted model. If this test is significant, DIF is assumed and an iterative Wald test procedure is performed to identify DIF items starting from a fully invariant model. Refinement of DIF is processed each time an item is flagged with DIF.

The performances were compared in terms of false detection (no simulated DIF), correct detection (simulated DIF), quality of DIF detection (affected groups and items) and group effect bias.

Results Rates of false DIF detection were low and ranged between 0.2 % and 1.6 %, and 1.0 % and 3.2 % for the anchor selection method and the DIF items identification method, respectively. DIF was correctly detected in 5 % to 99% and 24% to 98% of the cases for the anchor selection method and the DIF items identification method, respectively. Rates of correct detection increased with the number of groups affected by DIF, sample size and number of response categories.

Conclusion The DIF items identification method performed generally better than the anchor selection method. A sample size of 300 patients per group is required to achieve 80% of correct DIF detection which may limit the applicability of these methods, derived from educational sciences, in health sciences.

2: Linkage of HIV Treatment and Population-Based Surveillance Records in Rural South Africa

Dickman Gareta^{1,2,3,4}, Evelyn Lauren⁵, Khumbo Shumba⁶, Cornelius Nattey⁶, William Macleod^{6,7}, Matthew P. Fox^{6,7,8}, Koleka Mlisana^{4,10}, Matthias Egger^{2,11,12}, Dorina Onoya⁶, Kobus Herbst^{1,9}, Jacob Bor^{6,7}

¹Africa Health Research Institute, South Africa

²Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland

³Graduate School for Health Sciences, University of Bern, Bern, Switzerland

⁴School of Laboratory Medicine and Medical Sciences, University of KwaZulu Natal, Durban, South Africa

⁵Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA

⁶Health Economics and Epidemiology Research Office, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

⁷Boston University School of Public Health, Department of Global Health, Boston, MA, United States

⁸Department of Epidemiology, Boston University School of Public Health, Boston, MA, USA

⁹DSTI-SAMRC South African Population Research Infrastructure Network (SAPRIN), Durban, South Africa

¹⁰National Institute for Communicable Diseases, Johannesburg, South Africa

¹¹Centre for Infectious Disease Epidemiology and Research, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa

¹²Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK

Background Integrating HIV clinical records and demographic surveillance offers research opportunities to understand uptake of health services and healthcare quality and inequality

and improve patient outcomes. We implemented a graph-based record linkage algorithm to deduplicate and link HIV treatment records with population-based clinical and surveillance records in an HIV-endemic setting in rural South Africa.

Methods We deduplicated and linked data from four data sources: Africa Health Research Institute (AHRI) Health and Demographic Surveillance System (HDSS), AHRI Clinic and Hospital Information System (AHRILink), National Health Laboratory Service (NHLs), and Three Integrated Electronic Registers (TIER.Net, HIV care and treatment records). Data were collected between January 1, 2000, and July 31, 2024, through repeated HDSS surveys of over 22,000 households residing in AHRI's surveillance area and from one hospital and 17 clinics in Hlabisa sub-district, KwaZulu-Natal. The databases contained identifying attributes such as first name, surname, date of birth, gender, health facility, and South African national identity (ID) number, although typographical errors were common. We implemented a graph-based record linkage algorithm adapted from the Fellegi-Sunter model. The algorithm was trained and validated using a subset of records that contained valid national ID numbers. We assessed the algorithm performance by computing sensitivity, positive predictive value (PPV), and F-score, and computed descriptive statistics for different cohorts constructed from the linked database.

Results Deduplication and linkage of the four databases yielded a sensitivity of 91.7% and PPV of 94.8% (F-score= 0.932). Of 246,945 unique individuals from the HDSS, 43,325(17.54%) were HIV positive based on the data from the four data sources. Of these, 31,051(71.9%) had a record in TIER.Net or NHLs and 25,175(81.1%) had a record in TIER.Net. Of 64,140 unique individuals from TIER.Net, 25,175 (39.0%) individuals were household members in the HDSS. 16,074 (25.0%) of the individuals in TIER.Net had at least one hospital admission.

Conclusion Records from multiple data sources can be deduplicated and linked with a high degree of accuracy in a resource-poor HIV-endemic region in rural South Africa. This study paves the way for further advancements in clinical and population data integration, offering the potential to deepen our understanding of HIV epidemiology in a well-described population with a high prevalence of infectious and non-communicable diseases.

3: Investigating Statistical Inference for Consistency, Heterogeneity and Efficacy of Federated Learning Models: Insights from a Mega-Simulation of Real-World Data

Narayan Sharma¹, Gonzalo Durán-Pacheco¹, Jacek Chmiel², Eric Boernert¹, Doug Kelkhoff³, Vittorio P. Illiano¹, Gabriele Zilorri¹, Matthias Antonin¹, Bjoern Tackenberg¹, Dominik Heinzmann¹

¹F. Hoffmann - La Roche Ltd, Basel, Switzerland

²Avenga, Poland

³Hoffmann-La Roche Limited, Canada

Background Federated learning and analytics with medical data has emerged as a key solution to collaboratively train or fit machine-learning and statistical models to distributed datasets at different hospitals without compromising data privacy. Heterogeneity and site level differences necessitate a profound understanding of the operational characteristics of federated algorithms to ensure accuracy and interpretability of the models once they are applied to distributed real world medical datasets (RWD).

Different federated algorithms including GLMs and marginal structural models have been developed for an upcoming RWD analysis of persistence of treatment and its impact on patient outcome across hospitals on different continents. Their operational characteristics and accuracy has been investigated in a holistic simulation study and learnings will inform interpretability of the model outcomes once applied to RWD.

Methods We simulated RWD from three hospitals across 36 scenarios including heterogeneity (high, low), treatment effect (no, moderate and large), data missingness (no, at-random, not at random) and independent and identically distributed (IID) data process (IID, no-IID). A registry was used to estimate between hospital variability and other metrics to ensure simulation is as realistic as possible.

The workflow was set-up to generate data, model and extract summary statistics. We execute parallelly local and federated analysis for conditional and marginal models (linear, logistic and cox regression) for continuous, binary and time-to-event endpoints. We repeat each scenario for maximum 100 times (i.e., total 21,600 fitted models). Federated models were executed with DataSHIELD, an open source solution.

In addition, simulated hospital individual data were synthesized by a meta-analysis to compare it to the federated results.

Results The coefficients and standard errors for federated models and central models were highly similar, accurate to five decimal places, demonstrating that identical results can be obtained without centralizing data across hospitals. Under IID conditions, both federated and central models yielded unbiased results, while non-IID scenarios introduced biases with ground truth, particularly with increasing treatment effects. Notably, meta-analysis results strongly diverged, revealing an average hazard decrease of 1.2% to 6.3% under IID and an increase of 1.3% to 10.4% under non-IID for Cox-model, compared with the federated results across different scenarios.

Conclusion Federated models are appropriate for analysis of RWD from hospitals across

different countries and continents where privacy laws and other considerations restrict pooling (i.e. centralizing) of the data. For many scenarios, a simple meta-analysis could be misleading, favoring the federated approach.

4: Characterizing Medication Timelines in Huntington's Disease: A Cluster-Based Analysis of Treatment Patterns

Marc Dibling¹, Alexandra Durr², Sophie Tezenas du Montcel¹

¹Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, Paris, France

²Sorbonne Université, Paris Brain Institute - ICM, Inserm, CNRS, APHP, Hopital de la Pitié-Salpêtrière, Paris, France

Introduction Huntington disease (HD) is a rare inherited neurodegenerative disorder characterized by motor, cognitive, and psychiatric symptoms managed through diverse medication regimen strategies. However, the heterogeneity of clinical symptoms, adherence and tolerance poses significant challenges in determining the optimal care trajectory. Leveraging a national health claims database, this study aims to investigate HD patients' medication timeline to characterize the distinct strategies and assess heterogeneity factors.

Methods We identified 1,776 incident HD patients in the French health claims database (SNDS) between 2019 and 2023 and retrieved their tetrabenazine, antipsychotics and antidepressants pharmacy deliveries between 2009 and 2023. We then applied a three-steps data processing pipeline designed to detect patterns of medication use and enable inter-group comparisons. We first summarized individual patient treatment timelines via meta features (i.e. number of event occurrences, time gaps between occurrences) to secondly perform k-medoids timeline clustering using a large number of medoids (silhouette score > 0.3) to ensure that only highly similar timelines are grouped together thereby reducing complexity by selecting representative ones while preserving key patterns. Third, we performed hierarchical aggregative clustering of medoids to cluster similar timeline sequences and facilitate visualization of patterns of medication use. Number of clusters was selected to maximize the silhouette score. Finally, after the three steps data processing, we conducted an explanatory analysis to assess the relationship between treatment sequence clusters and baseline patient characteristics using ANOVA and chi-squared tests.

Results Mean age at detection of first HD diagnosis code was 57.5 [56.7-58.3] years with a 0.83 male-female ratio and most patients (99%) had at least one delivery of the drugs of interest between 2019 and 2023. We identified 7 clusters of medication timelines: one

with timelines indicating no treatment adherence or delivery, one for the individual use of each drug of interest, one for the combined use of antidepressants with antipsychotics and another for antidepressants with tetrabenazine. The last cluster included patients with a delayed treatment initiation. Baseline characteristics comparison indicated that patients not adhering to any treatment were significantly older, patients with regular tetrabenazine deliveries were significantly less hospitalized and there were significantly more men included in the cluster indicating a delayed treatment initiation.

Conclusion This study demonstrates the value of leveraging large registry data to analyze complex medication trajectories in Huntington disease. We identified distinct treatment patterns and adherence variations that will allow to understand disease management strategies on a national level.

5: Testing the Similarity of Healthcare Pathways Based on Transition Probabilities - A New Bootstrap Procedure

Zoe Lange¹, Holger Dette¹, Maryam Farhadizadeh², Nadine Binder²

¹Ruhr-University Bochum, Germany

²University Freiburg, Germany

Background Establishing a common standard of care within or across clinics and finding the best treatment strategies for diseases are important goals in the healthcare system. To contribute to achieving these goals we study the healthcare pathways of patients, consisting of sequences of diagnoses, treatment procedures, or hospital readmissions observed over time. Working with healthcare pathway data is attractive since this data is collected by clinics routinely and therefore, has a high availability. However, the healthcare pathways of different patients tend to be highly heterogeneous, even for common diseases. With our newly developed similarity testing approach, presented here, we can find patterns, namely typical pathways, in this heterogeneous data.

Methods We model the healthcare trajectory for a group of patients by a multistate model, where diagnoses, treatments or readmissions are considered as states that patients can transition to. We define the similarity of two multistate models in terms of the probabilities of a patient transitioning between states. This modelling of similarity enables us to formulate a similarity hypothesis test. If for two multistate models the difference between their transition probabilities is large, they are considered non-similar and fulfil the null hypothesis. If the difference is sufficiently small, they are considered similar and fulfil the alternative hypothesis. Groups with similar healthcare pathways, according to the test, are pooled into one

group, representing a typical pathway. Based on these pooled data sets, one can perform further estimation tasks like estimating the risks or probabilities for hospital readmission. The increased sample size, that results from pooling similar pathways, yields to more accurate statistical inference, especially in small sample settings as with heterogeneous pathways.

Results We introduce a special parametric bootstrap test that is tailored to our similarity hypotheses. We proof the validity of this test and investigate its performance in a comprehensive simulations study with different sample sizes, censoring rates, and similarity thresholds. Furthermore, we show how the results are applicable by discussing an example of prostate cancer data.

Conclusion Testing the similarity of seemingly heterogeneous healthcare pathways to identify typical pathways is a new and promising approach that accounts for small sample sizes and draws on available routine data. Beyond this, the presented bootstrap test can easily be adapted to other settings making it an attractive tool for general similarity testing problems, especially when only limited data is available.

Abstracts of Contributed Posters

Monday Posters at Biozentrum

Monday, 2025-08-25 10:45 - 11:30, Biozentrum, 2nd floor

1: Matching-Adjusted Indirect Comparison of Endoscopic and Craniofacial Resection for the Treatment of Sinonasal Cancer Invading the Skull Base

Florian Chatelet^{1,2,3}, Sylvie Chevret^{1,2}, MUSES collaborative group^{3,4,5}, Philippe Herman^{1,3}, Benjamin Verillaud^{1,3}

¹Université Paris Cité, France

²SBIM Hôpital Saint Louis APHP Paris, ECSTRA team

³ENT department Hôpital Lariboisière APHP Paris

⁴"ASST Spedali Civili di Brescia," University of Brescia, Brescia, Italy;

⁵"Ospedale di Circolo e Fondazione Macchi," University of Insubria, Varese, Italy

Background In surgical oncology, new techniques often replaces established methods without direct comparative studies, making it difficult to assess their actual effectiveness. This is particularly relevant for endoscopic endonasal approaches (EEA), which have progressively supplanted craniofacial resection (CFR) for sinonasal cancers invading the skull base. As a result, contemporary CFR-treated cohorts have become too small for direct comparisons, and randomised trials remain unfeasible due to ethical and logistical constraints. Matching-adjusted indirect comparison (MAIC) offers a statistical method to indirectly compare a contemporary individual-patient dataset (EEA) with a historical aggregate dataset (CFR), adjusting for confounding variables.

Methods We conducted a MAIC using individual patient data (IPD) from the MUSES cohort (EEA-treated patients) and aggregated data from Ganly et al. historical CFR cohort, including patients with skull base invasion. Key prognostic variables—including age, tumour histology, orbital and brain invasion, prior radiotherapy or surgery—were used to weight the MUSES cohort to match the CFR cohort.

Primary and secondary endpoints included overall survival (OS), recurrence-free survival

(RFS), perioperative mortality, surgical margins, and complication rates. Survival analyses were conducted using Kaplan-Meier estimations, log-rank tests, and Cox proportional hazards models, with bootstrap resampling for confidence interval estimation.

Results A total of 724 EEA-treated and 334 CFR-treated patients were analysed. Before MAIC, EEA was associated with significantly improved OS ($HR= 2.33$, 95% CI= 1.88–2.87, $p < 0.001$), and this benefit persisted after adjustment ($HR= 1.93$, 95%CI= 1.60–2.34, $p < 0.001$). RFS was initially higher in the EEA cohort ($HR= 1.39$, 95%CI= 1.14–1.69, $p = 0.001$) but was no longer statistically significant after adjustment ($HR= 1.06$, 95%CI= 0.91–1.23, $p = 0.63$). Perioperative mortality and complications were significantly lower in the EEA cohort compared to CFR. Clear resection margins were achieved in 79% of EEA cases and 71% of CFR cases ($OR= 0.67$, 95%CI= 0.50–0.90, $p = 0.008$), but this difference was no longer significant after MAIC adjustment ($OR= 1.15$, 95%CI= 0.93–1.40, $p = 0.36$).

Conclusion This study highlights the potential utility and limitations of MAIC in addressing selection biases in non-randomised comparisons. OS remained superior in the EEA group after adjustment, while RFS was similar between EEA and CFR. Perioperative mortality and complications were significantly higher with CFR, although both techniques achieved similar resection margin rates after adjustment. These findings support endoscopic surgery as a first-line approach for sinonasal cancers invading the skull base, provided it is technically feasible and performed in expert centres.

2: Information Borrowing in Phase II Randomized Dose-Ranging Clinical Trials in Oncology

Guillaume Mulier^{1,2}, Vincent Lévy³, Lucie Biard^{1,2}

¹Inserm U1342, team ECSTRRA. Saint-Louis Research Institute, Paris, France

²APHP, Department of Biostatistics and Medical Information, Saint-Louis hospital, Paris, France

³APHP, Clinical research department, Avicenne hospital, Paris, France

Introduction Over the past decades, the emergence of therapeutics such as immunotherapies and targeted therapies has challenged conventional trial designs, particularly single-arm studies. Selecting a single dose from phase I trials with limited follow-up, typically based solely on toxicity endpoints, has often resulted in suboptimal drug dosages. As a result, dose optimization in oncology is now encouraged by international initiatives such as the FDA's Project Optimus, the Optimal Cancer Care Alliance, and the Patient-Centered Dosing Initiative. This study was motivated by the case of Ibrutinib in chronic lymphocytic leukemia,

where the initially approved dose of 420 mg/day—determined through conventional phase I designs based on the maximum tolerated dose—was later found to achieve comparable response rates at lower doses. This highlights the potential value of dose-ranging phase II studies in oncology.

Assuming that borrowing information across doses can enhance statistical power, our objective is to compare various strategies for information borrowing in phase II randomized trials involving multiple doses of the same drug.

Methods The backbone phase II design considered is the Bayesian Optimal Design (BOP2), adapted for multi-arm settings with co-primary binary endpoints and interim analyses. This design employs a multinomial conjugate distribution within a Bayesian framework, with decision rules for stopping due to futility and/or toxicity based on posterior probabilities.

We adapted and compared different information borrowing approaches for estimating efficacy and toxicity:

- (i) power prior,
- (ii) incorporation of information from stopped arms,
- (iii) Bayesian hierarchical modeling,
- (iv) Bayesian logistic regression.

These methods were applied alongside BOP2 decision rules. A simulation study was conducted to assess the operating characteristics of each approach in a hypothetical randomized dose-ranging trial, evaluating efficacy and toxicity against reference values.

Results Our findings indicate that power prior, when applied without dynamic adaptation, is unsuitable as it increases false positive rates. Bayesian hierarchical modeling shrinks estimates toward a common mean, reducing variance but also inflating false positive rates. In contrast, Bayesian logistic regression provides a balanced trade-off, enhancing power to some extent while maintaining a lower false positive rate.

Conclusion Bayesian logistic regression, modeling both dose-toxicity and dose-efficacy relationships, combined with BOP2 decision rules, offers a promising approach for borrowing information in dose-ranging studies with a limited number of doses. However, designs without information borrowing provide stricter false positive control and should also be considered.

3: Information Borrowing in Bayesian Clinical Trials: Choice of Tuning Parameters for the Robust Mixture Prior

Vivienn Weru¹, Annette Kopp-Schneider¹, Manuel Wiesenfarth³, Sebastian Weber², Silvia Calderazzo¹

¹German Cancer Research Center (DKFZ), Germany

²Novartis Pharma AG, 4002 Basel, Switzerland

³Cogitars GmbH, Heidelberg, Germany

Introduction Borrowing external data for use in a current study has emerged as an attractive research area with potential to make current studies more efficient especially where recruitment of patients is difficult.

Methods Bayesian methods provide a natural approach to incorporate external data via specification of informative prior distributions. Potential heterogeneity between external and current trial data, however, poses a significant challenge in this context. We focus on the robust mixture prior, a convex combination of an informative prior with a robustifying component, that allows to borrow most when the current and external data are observed to be similar and least otherwise. This prior requires the choice of three additional quantities: the mixture weight, and the mean and dispersion of the robust component. Some choices of these quantities may, however, lead to undesirable operating characteristics. We systematically investigate this impact across combinations of robust component parameters and weight choices in one-arm and hybrid-control trials, where in the latter, current control data is informed by external control data. An alternative functional form for the robust component is also investigated.

Results For some parameter choices, losses may be still unbounded despite the use of dynamic borrowing for both testing and estimation, i.e. Type I error (TIE) rate may approach 1 while MSE may increase unconstrained. In the hybrid-control setting, the parameter choices further impact the size and shift of the “sweet spot”, where control of TIE rate and gain in power is observed. We observe that for such a sweet spot, the width negatively correlates with the maximum power gain. We further explore behavior of the mixture prior when adopting a heavy tailed distribution for the robust component, which is able to cap TIE rate and MSE inflation.

Conclusion The choice of the parameters of the robust component of the mixture prior as well as the mixture weights is non-trivial. All three parameter choices are influential, acting together and therefore their impact needs to be assessed jointly. We provide recommendations for these choices as well as considerations to keep in mind when evaluating operating characteristics.

4: A Bayesian Approach to Decision Making in Early Development Clinical Trials : An R Solution.

Audrey Te-ying Yeo

Independant

Early clinical trials play a critical role in Oncology drug development. The main purpose of early trials is to determine whether a novel treatment demonstrates sufficient safety and efficacy signals to warrant further investment (Lee & Liu, 2008). The new open source R package phase1b (Yeo et al, 2024) is a flexible toolkit that calculates many properties to this end, especially in the oncology therapeutic area. The primary focus of this package is on binary endpoints. The benefit of a Bayesian approach is the possibility to account for prior data (Thall & Simon, 1994) in that a new drug may have shown some signals of efficacy owing to its proposed mode of action, or similar activity based on prior data. The concept of the phase1b package is to evaluate the posterior probability that the response rate with a novel drug is better than with the current standard of care treatment in early phase trials such as Phase I. The phase1b package provides a facility for early development study teams to decide on further development of a drug either through designing for phase 2 or 3, or expanding current cohorts. The prior distribution can incorporate any previous data via mixtures of beta distributions. Furthermore, based on an assumed true response rate if the novel drug was administered in the wider population, the package calculates the frequentist probability that a current clinical trial would be stopped for efficacy or futility conditional on true values of the response, otherwise known as operating characteristics. The intended user is the early clinical trial statistician in the design and interim stage of their study and offers a flexible approach to setting priors and weighting.

5: Designing Clinical Trials in R with Rpact and crmPack

Daniel Sabanés Bové¹, Gernot Wassmer², Friedrich Pahlke²

¹RCONIS, Taiwan

²rpact GbR, Germany

The focus of this poster will be on clinical trial designs and their implementation in R. We will present rpact, which is a fully validated, open source, free-of-charge R package for the design and analysis of fixed sample size, group-sequential, and adaptive trials. We will summarize and showcase the functionality of rpact.

In addition, we will also briefly present *crmPack*, which is an open source, free-of-charge R package for the design and analysis of dose escalation trials.

Together, *ract* and *crmPack* enable the implementation of a very wide range of clinical trials. The poster presentation aims to increase the visibility of the two open source packages in the clinical biostatistics community, and allow for discussions about future developments.

6: Leveraging on Historical Controls in the Design and Analysis of Phase II Clinical Trials

Zhaojin Chen¹, Ross Andrew Soo^{2,3}, Bee Choo Tai^{1,4}

¹Saw Swee Hock School of Public Health, National University of Singapore, Singapore

²Department of Haematology-Oncology, National University Cancer Institute Singapore, Singapore

³Cancer Science Institute of Singapore, National University of Singapore, Singapore

⁴Yong Loo Lin School of Medicine, National University of Singapore, Singapore

Background In oncology, phase II trials are commonly used to screen novel agents for solid tumour by following a single-arm design. All patients receive a concurrent treatment (CT) and their overall objective response rate is compared with some pre-defined threshold. However, evidence has suggested that such a design often results in false claims of efficacy. This not only causes waste in time and resources but is also of great ethical concern for trial participants. This study thus aims to improve the current design by incorporating a historical control (HC) arm for more appropriate treatment evaluation.

Methods For treatment evaluation using HCs, major challenges involve imbalance in baseline characteristics, unmeasured baseline variables and temporal drift of disease outcomes. To tackle these problems, we adopted three main statistical approaches, namely regression adjustment (RA), inverse probability of treatment weighting (IPTW_PS) and matching (MC_PS) based on propensity score, to reduce potential confounding bias when evaluating the effect of treatment. Simulation studies were conducted for null, small, moderate and large treatment effect based on a binary disease outcome, assuming sample sizes of 100 and 200 with equal treatment allocation. Bias, mean squared error (MSE), coverage probability, type I error and power were used to evaluate their performances. These methods were then applied to the PLASMA phase II trial using HCs from the previously completed AURA 3 phase III trial.

Results Simulation results showed that the RA method slightly overestimates, whereas the

IPTW_PS method slightly underestimates treatment effect as it goes from null to large. Bias of the MC_PS method can be in either direction and reduces in magnitude when more HCs are available. As the level of imbalance in baseline characteristics increases, the bias and MSE increase and power decreases. All three methods are sensitive to unmeasured baseline confounders, but the RA method appears to be more sensitive to model misspecification as compared to the propensity score based methods.

Conclusion Consistent with existing literature, our study found that phase II trials incorporating HCs should be recommended for diseases with well-known mechanisms. Moreover, when there are a large number of HCs available, the MC_PS generally performs better than the other two methods with desirable bias, MSE, type I error and power.

7: Design of a Research Project to Evaluate the Statistical Utility after Transformation of a CDISC Database into OMOP Format

Claire Castagné¹, Amélie Lambert¹, Jacek Chmiel², Alberto Labarga³, Eric Boernert³, Lukasz Kaczmarek³, Francois Margraff³, David Pau¹, Camille Bachot¹, Thomas Stone³, Dimitar Toshev³

¹Roche, France

²Avenga, Germany

³F. Hoffmann-La Roche AG

Interoperability between databases is an important issue, to facilitate analyses from multiple sources. The OMOP (Observational Medical Outcomes Partnership) format is increasingly used in Europe, particularly in France. After a targeted bibliographical review of data sources and standard formats used, no article precisely assesses the loss of data and/or information following transformation to the OMOP format. The aim of this work is to assess the statistical and scientific usefulness of the OMOP format.

An observational study in early breast cancer was conducted in 2019. The database is currently in CDISC SDTM format.

The first step of the project involves transforming the SDTM database into OMOP format.

In the second step, a statistical analysis of the data in OMOP format will be carried out.

In the third step, all the results will be compared with the initial results, using quality indicators to assess the loss of information:

- indicators regarding transformation to OMOP format, such as the number of observations or variables not transformed,
- indicators regarding the number of statistical tables not generated,
- indicators regarding the reliability (no loss of information, partial loss, complete loss) of results obtained by comparing SDTM vs OMOP results

315 patients were included in the study, the database structure is made of 7 CDISC domains containing 73 variables: 25 continuous and 48 categorical regarding patient, disease, surgery and treatments characteristics.

Age at treatment initiation was 52.2 (11.8) years, distribution of SBR grade evaluating disease severity was: grade III 50.7% of patients, II 45.7% and I 1.6%.

40.3% of patients met the primary outcome, evaluated at surgery by the pathological complete response.

OMOP database will start in February 2025 and results will be available at the congress: descriptive analyses (univariate, bivariate), correlation matrices, modeling and survival analysis will first be performed on the raw study data (SDTM format), then these same analyses will be reproduced on the OMOP datasets. The usual statistical indicators (percentage of missing data, data dispersion, etc.) and the maintenance of relationships between variables will be used to quantify the differences observed between the databases in the different formats.

This work will make it possible to assess the statistical usefulness remaining after the switch to OMOP format, thanks to a synthesis of indicators, and to ensure the reproducibility of classic statistical analyses.

At the conference, the results/indicators observed on the OMOP format database will be presented and discussed in relation to the initial results.

8: Introducing CAMIS: An Open-Source, Community Endeavor for Comparing Analysis Method Implementations in Software

Yannick Vandendijck^{1,4}, Christina Fillmore^{2,4}, Lyn Taylor^{3,4}

¹J&J Innovative Medicine, Belgium

²GSK, UK

³Parexel, UK

⁴on behalf of the CAMIS working group

Try this in R: > round(2.5), and it will give the result of 2.

Try this in SAS: > data rounding; x = round(2.5); run; and it will give the result of 3.

Seriously?

Introduction Statisticians using multiple statistical software (SAS, R, Python) will have found differences in analysis results that warrant further exploration and justification. These possible discrepancies across statistical software for a similar analysis can cause unease when submitting these results to a regulatory agency, as it is uncertain if the agency will view these differences as problematic. This becomes increasingly important since the pharma industry is more and more turning to open-source software like R to handle complex data analysis, drawn by its flexibility, innovation, added value and cost-effectiveness.

Knowing the reasons for differences (different methods, options, algorithms, etc.) and understanding how to mimic analysis results across software is critical to the modern statistician and subsequent regulatory submissions.

CAMIS:

This talk will introduce the PHUSE DVOST CAMIS (Comparing Analysis Method Implementations in Software) project. The aim of CAMIS is to investigate and document differences and similarities between different statistical software (SAS, R, Python) to help ease the transitions to new languages by providing comparison and comprehensive explanations. CAMIS contributes to the confidence in reliability of open-source software by understanding how analysis results can be matched perfectly or knowing the source of any discrepancies.

In this talk, I will discuss the objectives of the CAMIS project, identify some key results on differences and similarities between SAS and R, show how we collaborate on CAMIS across companies/ industries/ universities in the open-source community.

Conclusion In the transition from proprietary to open-source technology in the industry, CAMIS can serve as a guidebook to navigate this process.

<https://psiaims.github.io/CAMIS/>

<https://github.com/PSIAIMS/CAMIS>

9: Assessing Covariates Influence on Cure Probability in Mixture Cure Models using Martingale Difference Correlation

Blanca E. Monroy-Castillo, M. Amalia Jácome, Ricardo Cao

Universidade da Coruña, Spain

Background Cure models analyze time-to-event data while accounting for a subgroup of individuals who will never experience the event. A fundamental question in these models is whether the cure probability is influenced by specific covariates. However, formal statistical tests for assessing covariate effects remain limited. Martingale difference correlation (MDC) provides a non-parametric measure of dependence, where $MDC(Y|X) = 0$ if and only if $E(Y|X) = E(Y)$, meaning X has no impact on the expectation of Y . This makes MDC a promising tool for testing covariate effects on cure probability.

Methods We propose a non-parametric hypothesis test based on MDC to evaluate the effect of covariates on the cure probability. A key challenge is that the cure indicator (ν) is only partially observed due to censoring. To address this, we estimate the cure status before applying the test. The methodology is validated through extensive simulation studies, assessing its power and robustness under different scenarios. Additionally, we apply the proposed test to data from a randomized clinical trial on rheumatoid arthritis treatment to identify covariates influencing disease remission.

Results Simulation studies demonstrate the effectiveness of the proposed method in detecting covariate effects on the cure probability. When applied to the clinical trial data, the test identifies specific covariates associated with an increased probability of experiencing a flare-up. These findings provide new insights into factors influencing disease progression and treatment response in rheumatoid arthritis patients.

10: Aligning Estimators to Treatment Effects in the Presence of Intercurrent Events in the Analyses of Safety Outcomes

Pedro Lopez-Romero¹, Brenda Crowe², Philip He³, Natalia Kan-Dobrosky⁴, Andreas Sashegyi², Jonathan Siegel⁵

¹Novartis, Spain

²Eli Lilly, USA

³Daiichi Sankyo Inc, USA

⁴AbbVie Inc, USA

⁵Bayer, USA

Introduction The evaluation of safety is a crucial aspect of drug development. The ICH Estimand Framework (EF) defines clinically relevant treatment effects in the presence of intercurrent events (ICE) and can enhance this evaluation. However, its application in safety evaluation is uncommon. Additionally, sometimes it is not evident which specific estimand a given estimator is targeting, leading to the implementation of analytical strategies that may not align with the treatment effect of clinical interest.

Methods This work reviews the clinical questions or treatment effects (estimands) that are most common in the safety evaluation of drugs and the strategies outlined in the EF that reflect those treatment effects. We examine the most common statistical estimators used to assess the risk of drugs, including incidence proportions, Aalen-Johansen estimator, expected adjusted incidence rates and 1 minus Kaplan-Meier, focusing on the interpretation of the estimates and on the estimand they target, depending on how ICEs are defined for analysis, e.g. ignored, censored, or as competing events. By understanding a) the treatment effects that we can feasibly define in the presence of ICEs and b) the estimand that is targeted by different estimators, our goal is to define treatment effects that are clinically meaningful for the evaluation of safety, and to use the estimator that aligns with the treatment effect of interest, so that the treatment effect estimates are meaningful and interpretable.

Results Our review includes treatment effects or estimands that are relevant to the evaluation of drug safety, such as treatment policy, hypothetical and while-on-treatment, considering ICE such as early treatment discontinuation or use of rescue medication. We explain why the common estimators target different estimands, helping researchers to select the estimator that aligns with the treatment effect of interest. A misalignment between the estimator and the treatment effect of interest can eventually lead to misinterpretations of safety results that potentially can compromise the understanding about the safety profile of a drug.

Conclusions Applying the EF to safety evaluation can improve the interpretability of treatment effects in clinical development, both in the area of signal detection and in the analysis of selected adverse events of special interest. By clearly defining the estimand and selecting the appropriate statistical method, researchers can ensure that their analyses align with clinically relevant questions. This approach enhances the accuracy and reliability of safety assessments, ultimately contributing to better-informed decision-making in drug development by regulators, physicians, patients and other stakeholders.

11: CUtools: An R Package for Clinical Utility Analysis of Predictive Models

María Escorihuela Sahún¹, Luis Marianos Esteban Escaño¹, Gerardo Sanz², Ángel Borque-Fernando³

¹Department of Applied Mathematics, Escuela Universitaria Politécnica La Almunia, University of Zaragoza, Spain

²Department of Statistical Methods, University of Zaragoza, Spain

³Urology department, Miguel Servet university hospital, Spain

This work presents a new library in R that provides statistical techniques to validate and evaluate a prediction model both analytically and graphically. The library offers the functions CUC_plot, CUC_table, Efficacy, Efficacy_curve, and Efficacy_test to construct the clinical utility curve, a table of clinical utility values, the efficacy of a biomarker, the efficacy curve, and a test to compare the efficacy of biomarkers.

The purpose of predictive models in clinical diagnosis is to define a biomarker that accurately predicts the occurrence of an event related to a disease. To analyse the predictive capability of a biomarker, this library provides, as an initial output, the clinical utility curve via the CUC_plot function. Clinical utility assesses the benefit of a biomarker used as a dichotomous classifier with a cut-off point. On the X-axis, the possible cut-off points of a biomarker as a continuous variable are plotted, and on the Y-axis, two magnitudes appear: the percentage of misclassified events and the percentage of individuals below the cut-off point. These values represent the false negative rate and the number of treatments avoided when applying the model. Additionally, the CUC_table function provides the numerical values represented graphically.

Another way to analyse the clinical utility of a biomarker is by calculating its efficacy. To study efficacy, this library offers an analytical result with the Efficacy function and a graphical result with the Efficacy_curve function. On the one hand, the numerical value of the marker's efficacy is obtained as the difference between the treatments avoided by the model and the misclassified events; on the other hand, a graph is produced in which the X-axis shows the values of misclassified events versus the efficacy of the proposed model.

12: Impact of Particulate Matter 2.5 Levels on Chronic Obstructive Pulmonary Disease: An Analysis of Nationwide Claims Data in Thailand

Pawin Numthavaj¹, Tint Lwin Win¹, Chaiyawat Suppasilp¹, Wanchana Ponthongmak¹, Panu Looareesuwan¹, Suparee Boonmanunt¹, Oraluck Pattanaprateep¹, Prapaporn Pornsuriyasak¹, Chathaya Wongrathanandha², Kriengsak Vareesangthip³, Phunchai Charatcharoenwitthaya⁴, Atiporn Ingsathit¹, Ammarin Thakkinstian¹

¹Department of Clinical Epidemiology and Biostatistics, Faculty of Medicine Ramathibodi Hospital, Mahidol University, Thailand

²Department of Community Medicine, Faculty of Medicine Ramathibodi Hospital, Mahidol University

³Division of Nephrology, Department of Medicine, Faculty of Medicine Siriraj Hospital, Mahidol University

⁴Division of Gastroenterology, Department of Medicine, Faculty of Medicine Siriraj Hospital, Mahidol University

Introduction Particulate matter 2.5 (PM 2.5) levels have been associated with morbidity and mortality in chronic obstructive pulmonary diseases (COPD). We explored the association between levels of PM 2.5 and exacerbations documented in a Thailand national database claim by the National Health Security Office, which covers about 70% of the Thai population.

Methods We extracted the data of COPD exacerbations from the identified international classification of disease - 10th version (ICD-10) among patients who were more than 40 years old, as well as verified the information upon documented procedures of usage of nebulizer, intubation, ventilator use, and temporary tracheostomy performed. Data of PM 2.5 levels were estimated from satellite data formula verified with ground datapoint collection. Incidences of COPD exacerbation were then calculated for each week of each district of provinces across Thailand and were modelled for a relationship with the exposure of PM 2.5 in the previous seven days, adjusted for age, gender, and baseline rates of diagnosed comorbidities of cancer, asthma, hypertension, heart failure, anxiety, depression, obesity, diabetes, and dyslipidaemia in mixed-effect Poisson regression with random intercept using R. We also explored the formula for averaging PM 2.5 in the area with the standard average and the area-weighted PM 2.5 level on the model fitness.

Results A total number of 407,866 verified COPD patients from January 2017 until December 2002 were identified, corresponding to 1,687,517 hospital visits. Among these visits, exacerbation or visits that required lower airway interventions happened in a total of 1,687,517 visits (9.9%). Multivariate Poisson regression analysis found that the incidence rate ratio (IRR) of COPD exacerbation of 1.00098 for each 1 microgram per cubic metre of PM 2.5 increment (95%CI 1.00091 – 1.00106). The weighted PM 2.5 formula was found

to have less Akaike information criterion and Bayesian information criterion values in the multivariate model compared to the standard average PM 2.5 calculation used in previous studies (6,278,282 vs. 6,279,384, and 6,279,627 vs. 6,278,539, respectively).

Conclusion From our analysis, the PM 2.5 level is associated with an increase in the occurrence of COPD exacerbation. We also found that the weighted formula used to calculate the exposure levels seems to fit the data more than the regular formula used in the traditional formula in the literature.

13: Changes in Health Services Use of a Cohort of COPD Patients from a Pre-Pandemic to a COVID-19 Pandemic Period

Jose M Quintana^{1,2,4,5}, **Maria J Legarreta**^{1,2,4,5}, **Nere Larrea**^{1,2,4,5}, **Irantzu Barrio**^{2,4,5,6},
Amaia Aramburu^{1,3}, **Cristóbal Esteban**^{1,3}

¹Osakidetza/SVS - Galdakao-Usansolo Hospital, Spain

²Instituto Biosistemak, Bilbao, Spain

³Instituto BioBizkaia, Barakaldo, Spain

⁴REDISSEC

⁵RICAPPS

⁶UPV/EHU

Background COVID-19 pandemic had negative effects on health especially in people with chronic diseases. We evaluate the differences in health services use among patients with chronic obstructive pulmonary disease (COPD) during the period of 2017-2019 compared to 2020-2022, COVID pandemic period.

Methods Cohort of patients recruited from different hospital who had an admission due to COPD exacerbation. Sociodemographic and clinical data were collected from all participants at 2016. A follow up was performed at 2022 with those who agreed to participate, focusing on their use of health services. This included number hospital admissions by any cause, to ICU, visits to Emergency Room, consultations with primary care physician, nurse, or medical specialists. The data was collected for the periods of 2017-2019 and 2020-2022. A sample of patients in the form of paired data was generated where time 1 corresponds to the years 2017-2019 and time 2 of the same patient corresponds to the 2020-2022 period. From these data, multivariate negative binomial regression models were developed for all the number of service usage even data with random effects for patients. Models were adjusted by study period age, Charlson Index, previous admissions and SARS-CoV-2 infection or hospital admission on period 2.

Results Out of the original cohort of 1,401 patients, 703 (50.2%) died during the follow-up period. Of the remaining, 314 (45%) chose not to participate in the study, while 384 (55%) did participate. The mean age of the participants was 69.2 years (SD: ± 9.8), with men constituting 72.1% of the sample. We observed a statistically significant reduction in the number of hospital admissions, ICU admissions, emergency visits, and face-to-face visits with primary care doctors from the first period to the second period. However, there was no significant change in the number of face-to-face consultations with primary care nurses or pneumologists. Having a SARS-CoV-2 infection or being admitted for it during the second period was associated with an increase in hospital admissions, emergency visits, and face-to-face consultations with pneumologists and primary care nurses. Additionally, SARS-CoV-2 infection influenced the face-to-face visits to primary care doctors, but neither factor affected ICU admissions.

Conclusion COVID-19 pandemic had an important negative effect on patients with COPD. On the one hand, access to the use of most health services in these patients decreased significantly. On the other hand, having had a SARS-CoV-2 infection or a hospital admission by it was related to a greater use of these health services.

14: The ISARIC Clinical Epidemiology Platform: Standardized Analytical Pipelines for Rapid Outbreak Response

Esteban Garcia-Gallo¹, Tom Edinburgh¹, Sara Duque¹, Leonardo Bastos², Igor Tona Peres², Elise Pesonel¹, Laura Merson¹

¹Pandemic Sciences Institute, University of Oxford (United Kingdom)

²Pontifical Catholic University of Rio de Janeiro (Brazil)

Background

The International Severe Acute Respiratory and Emerging Infection Consortium (ISARIC) is a global research network facilitating rapid clinical responses to infectious disease outbreaks. Comprising 60 members across 133 countries, ISARIC has generated critical evidence for diseases such as COVID-19, dengue, Ebola, and mpox. Its guiding principles—Prepare, Integrate, Collaborate, and Share—support research readiness, integration with public health systems, strong partnerships, and open-access resource-sharing.

Since 2012, the ISARIC Clinical Characterisation Protocol (CCP) has enabled standardized, adaptable investigations of high-consequence pathogens. During the COVID-19 pandemic, ISARIC's CRF was widely adopted, contributing to a dataset of one million patients. Lessons from past outbreaks underscore the need for both flexibility and standardization in clinical

research. A decentralized approach ensures local data ownership while enabling global integration, scalability, and equitable collaboration—key principles driving the development of the ISARIC Clinical Epidemiology Platform (ISARIC-CEP).

Methods

The ISARIC-CEP consists of three tools—ARC, BRIDGE, and VERTEX—designed to streamline data collection, curation, analysis, and evidence generation. ARC provides a machine-readable library of standardized CRF questions, BRIDGE automates CRF generation for seamless REDCap integration, and VERTEX is an open-source application comprising three packages:

1. **get_REDCap_Data:** Harmonizes and transforms ARC-formatted REDCap data into analysis-ready dataframes.
2. **ISARICAnalytics:** A set of Reusable Analytical Pipelines (RAPs) standardizing key epidemiological analyses, including descriptive statistics, data imputation, regression models, feature selection, and survival analysis.
3. **ISARICDraw:** Generates interactive dashboards with customizable outbreak-specific visualizations using Plotly.

VERTEX supports insight panels, organizing outputs into thematic sections, and its adaptable framework enables secure customized dashboards for multiple projects.

Results

The ISARIC-CEP has accelerated clinical research responses, including studies on dengue in Southeast Asia and Brazil. By providing openly accessible tools, it has facilitated high-quality analyses for both scientific and public health communities. Key resources include:

- **ARC:** <https://github.com/ISARICResearch/ARC>
- **BRIDGE:** <http://bridge.isaric.org/>
- **VERTEX:** <https://github.com/ISARICResearch/VERTEX>
- **Public Dashboard Example:** <http://vertex-observationalcohortstudy-mpox-drc.isaric.org>

Conclusions

The ISARIC-CEP accelerates outbreak research by ensuring that during an initial response, most time is spent on data capture, not on harmonization, curation, or preparation for analysis. VERTEX's RAPs streamline analyses, allowing standardized workflows to be shared and

adapted across outbreaks, reducing duplication and improving efficiency. Our goal is to build a collaborative community where researchers contribute RAPs, making validated methodologies easily integrable and reusable, amplifying their real-world impact. This approach strengthens clinical research sites, providing automated tools that enhance local capacity and ensure rapid, reproducible, and scalable outbreak analyses.

15: Topic Modelling and Time-Series Analysis to Explore Methodological Trend Evolution

Gabrielle Gauthier-Gagné, Tibor Schuster

McGill University, Canada

Background Statistical methodology used in biomedical research is evolving rapidly, driven by advances in biostatistical approaches and increased integration of machine learning techniques and causal inference frameworks. This convergence is reshaping the methodological foundations that underlie the analysis and interpretation of biomedical data in the literature. Both applied and methodological researchers may wish to explore these trends in their field to better understand the associated implications for evaluating, planning and conducting future studies. However, exploring these trends using conventional literature reviews is both time-consuming and requires periodical updates as the field develops. Therefore, we propose leveraging topic modelling and time-series analysis to explore methodological trend evolution which can easily be replicated and updated.

Methods We considered two parallel case studies to informally assess the utility of the proposed approach: examination of i) the literature on clinical trials and ii) literature pertaining to medical records. We employed readily available APIs to systematically extract PubMed abstract data related to studies which conducted clinical trials or examined medical records, respectively, in the last 10 years. Abstract text data was tokenized and structured as a document-term matrix (DTM). A large language model software was used to generate an exhaustive dictionary of terms (uni- and bigrams) commonly used in statistics and machine learning. The DTM was reduced to include only the terms corresponding to the entries of the derived term dictionary. Very common terms were additionally excluded. Latent Dirichlet Allocation was used to uncover latent topics across abstracts and to enable mapping of the distribution of topics within abstracts. Time-series analysis was used to characterize and visualize the trends of (average) topical prevalence over time (months), leveraging abstract publication dates and corresponding topic distributions.

Results The search identified 166, 932 and 7,999 unique abstracts relating to clinical trials

or medical record studies, respectively, for review. The generated statistical and machine-learning term lists contained 1803 statistical- and 200 machine-learning-related terms and bigrams. Both time series analyses and visualizations of topic trends over the past decade indicate dynamic and distinct shifts in the landscape of statistical methodology specific to each case study.

Conclusion We demonstrate that topic modelling paired with time-series analysis are powerful tools for methodological researchers to explore the evolution of statistical methodologies in their field over time.

16: Post-Stroke Facial Palsy: Prevalence on Admission, Risk Factors, and Recovery with Hyperacute Treatments

Zewen Lu^{1,2}, Havva Sumeyye Eroglu³, Halvor Næss⁴, Matthew Gittins^{1,2}, Amit K Kishore^{2,5}, Craig J Smith^{2,5}, Andy Vail^{1,2}, Claire Mitchell³

¹Centre for Biostatistics, University of Manchester, Manchester Academic Health Science Centre, UK

²Manchester Centre for Clinical Neuroscience, Geoffrey Jefferson Brain Research Centre, Manchester Academic Health Centre, Manchester Academic Health Centre, Salford Care Organisation, Northern Care Alliance NHS Foundation Trust, UK

³Division of Psychology, Communication & Human Neuroscience, Geoffrey Jefferson Brain Research Centre, University of Manchester, Manchester, UK

⁴Department of Neurology, University of Bergen, Haukeland University Hospital, Bergen, Norway

⁵Division of Cardiovascular Sciences, Faculty of Biology, Medicine and Health, University of Manchester, UK

Background

Facial palsy affects 40 - 50% of stroke survivors, impacting quality of life, communication, and emotional expression. This study estimated its prevalence, identified risk factors, assessed 7-day recovery post-admission and examined associations between hyper-acute treatments (intravenous thrombolysis [IVT] and mechanical thrombectomy [MT]) and recovery in acute ischaemic stroke (AIS) patients.

Methods

This was a retrospective individual data analysis of the Bergen NORSTROKE registry with 5987 patients (2006–2021). Only 2293 patients with facial palsy were included in our recovery analysis. We further investigated the association of hyper-acute treatments with facial palsy

recovery for 1954 patients with AIS. The complete case analysis was used in each stage of analysis due to minimal missing data. Facial palsy was assessed via the National Institute of Health Stroke Scale. Prevalence and severity of facial palsy on admission were analysed using descriptive statistics, while multifactorial logistic regression explored associations with demographics, stroke subtypes, and neurological symptom clusters. Kaplan-Meier survival curves estimated recovery rates within seven days of admission, and Cox proportional hazards models identified factors associated with recovery. The association between hyper-acute treatments and recovery was assessed using Cox models with time-dependent covariates, adjusting for baseline characteristics.

Results

Facial palsy was present in 43% of patients on admission, with 40% experiencing minor or partial paralysis and 3% complete paralysis. Significant risk factors included sex, age, admission motor and sensory function, and ischaemic stroke. By day 3, 25% of patients had recovered, but over 60% still had facial palsy by day 7. Better admission motor and sensory function were strongly associated with recovery. Receiving IVT showed a significant association with better recovery in unadjusted analyses, but neither IVT nor MT were significant in adjusted models.

Conclusions

Post-stroke facial palsy is common on admission, with less than 40% of patients recovering within the first week. This highlights the need for targeted monitoring and rehabilitation. Further research is to explore the role of hyper-acute treatments in longer-term recovery.

17: Evaluating Outlier Detection Methods in Real-World Growth Data: A Sensitivity Analysis of Imperfect Data in a Cluster Randomised Controlled Trial

Maryam Shojaei Shahrokhbadi¹, Mohadeseh Shojaei Shahrokhbadi², Bram Burger³, Ashley J. Adamson², Dawn Teare²

¹Hasselt University, Belgium

²Newcastle University, UK

³Uppsala University, Sweden

Background Growth studies with longitudinal measurements need outlier detection methods that can consider diverse, individual growth trajectories. Several methodological approaches have been developed, with distinct underlying assumptions, which can lead to differing results, potentially influencing study conclusions. To assess the reliability and robustness of primary analyses, we conducted a sensitivity analysis exploring the impact of multiple outlier detection

methods on findings from the MapMe 2 study [1].

Methods The MapMe 2 study, a cluster randomised controlled trial (cRCT), evaluated whether incorporating the MapMe 2 intervention into existing National Child Measurement Programme (NCMP) feedback letters improved child weight outcomes after one year. The primary outcome compared the change in BMI Z-score between intervention and control groups, including all children irrespective of baseline weight status, and specifically among children with a BMI Z-score > 1.33 at baseline. While the study initially used static WHO cut-offs to identify extreme or biologically implausible values (BIVs), in this large-scale trial we explored alternative outlier detection methods. Five approaches were compared to the original sBIV method [2]: (1) modified BIV detection (mBIV), (2) single-model outlier measurement detection (SMOM), (3) multi-model outlier measurement detection (MMOM), (4) multi-model outlier trajectory detection (MMOT), and (5) clustering-based outlier trajectory detection (COT). We then evaluated the impact of these methods on the study findings.

Results Different outlier detection methods resulted in variations in the number of subjects analysed and slight changes in the estimated effect of the MapMe 2 intervention on BMI Z-score change at one year. However, these differences were minimal, and the overall trends remained consistent.

Conclusion Sensitivity analyses under varying assumptions yielded results consistent with the primary analysis, confirming its robustness and reinforcing confidence in the trial findings.

References

1. Adamson AJ, et al. Can embedding the MapMe2 intervention in the National Child Measurement Programme lead to improved child weight outcomes at one year? 2021. Trial registration: [ISRCTN12378125]. Available from: <https://www.isrctn.com/ISRCTN12378125>.
2. Massara P, Asrar A, Bourdon C, Ngari M, Keown-Stoneman CD, Maguire JL, Birken CS, Berkley JA, Bandsma RH, Comelli EM. New approaches and technical considerations in detecting outlier measurements and trajectories in longitudinal children growth data. BMC Medical Research Methodology. 2023 Oct 13;23(1):232.

18: Latent Class Analysis on Intersectional Social Identities and Mental Wellbeing among Ethnic Minority Youth in Aotearoa New Zealand

Arier Lee¹, Shanthi Ameratunga^{1,2}, Rodrigo Ramalho¹, Rachel Simon-Kumar¹, Vartika Sharma¹, Renee Liang¹, Kristy Kang¹, Terryann Clark³, Terry Fleming⁴,

Roshini Peiris-John¹

¹School of Population Health, University of Auckland, Auckland, New Zealand

²Population Health Gain, Population Planning Funding and Outcomes Directorate, Te Whatu Ora – Health New Zealand, Auckland, New Zealand

³School of Nursing, University of Auckland, Auckland, New Zealand

⁴School of Health, Victoria University of Wellington, Wellington, New Zealand

Background / Introduction Ethnic minority youth in Aotearoa New Zealand who identify as Asian, Middle Eastern, Latin American, or African navigate multiple shifting identities. Conventional approaches in the literature often frame their experiences through a single social dimension, such as ethnicity. However, this limits deeper insights into how overlapping social identities, linked to broader structural inequities, affect emotional wellbeing. Using an intersectional framework, this study explored how multiple social identities and affiliations influence the mental health and wellbeing of ethnic minority young people.

Methods We analysed cross-sectional data from 2,111 ethnic minority youth (99% aged 13 to 19) who participated in a population-based secondary school survey in New Zealand in 2019. Latent Class Analysis (LCA) was employed to identify unobserved social affiliation groups based on categorical variables, including sex, sexual and gender identities, religion, perceived ethnicity, migrant generational status, disability, and material deprivation. LCA was also applied to nine family connectedness indicators (e.g., trust in sharing feelings with a family member), classifying participants into distinct family support groups. Multiple logistic regression models were used to predict the outcomes of mental health and wellbeing, by LCA-identified social affiliation and family support groups, and experiences of discrimination and bullying.

Results LCA identified four distinct social affiliation groups among ethnic minority youth:

1. Least marginalised, mixed migration generations
2. Some marginalised affiliations, mainly overseas-born
3. Some marginalised affiliations, mainly NZ-born
4. Multiply marginalised, mixed migration generations

The least marginalised group (Group 1) reported the best mental health and wellbeing outcomes, followed by Groups 2 and 3, while the multiply marginalised group (Group 4) exhibited the highest risks of adverse health outcomes. Independent of social affiliation group, experiences of discrimination and bullying were strongly associated with increased risks of poor

mental health. However, higher levels of family support significantly reduced these risks across all social affiliation groups.

Conclusion Marginalised social identities have cumulative harmful effects on the mental health and wellbeing of ethnic minority youth, but family support can serve to, mitigate some, but not all risk. The use of LCA enabled the classification of participants into distinct social affiliation groups based on multiple intersecting social identity variables, without assuming their independence, thus providing a more nuanced relationship between identity and mental health outcomes. These findings underscore the need to create inclusive and supportive environments for ethnic minority youth and their families.

19: Using Multiple Imputation in Real-Word Data Studies to Aid in the Identification of Predictors of Response While Addressing Missing Data

Jozefien Buyze¹, Lada Mitchell², Lorenzo Acciarri³

¹Johnson & Johnson, Beerse, Belgium

²Johnson & Johnson, Allschwil, Switzerland

³CRO Valos (J&J partner), Genova, Italy

Background In real-world data (RWD) studies, the inference drawn from estimates can be jeopardized by missingness in key variables. Recent guidance from the FDA (March 2024) and ICH EMA (May 2024) emphasizes the importance of addressing this issue. This research aims to address missing data for covariates in RWD studies. The hypothesis is that multiple imputation helps to reduce bias and improve validity, reliability, and efficiency of the estimation methods.

Methods Multiple imputation was applied, assuming data is missing at random (MAR). Multiple imputation preserves all cases and accounts for uncertainty due to missing data. It is crucial to recognize that if the MAR assumption is violated, the results may be biased. Given the non-monotone missing data pattern observed, we applied the fully conditional method for imputing missing variables. This method does not rely on joint distribution but generates separate conditional distributions for each variable needing imputation (Van Buuren 2007).

Understanding the effectiveness of the standard of care and the predictors for it remains an area of unmet need. The case study utilizes data pooled from two prospective single-arm oncology real-world data (RWD) studies, where missing data is present in several baseline covariates relevant to the statistical model for the effectiveness variable, *i.e.* for the overall

response rate (ORR).

The performance of multiple imputation in different scenarios with varying amounts of missingness was investigated via simulations.

Results The models were applied on pooled data (N=302) of two RWD studies. The results of the models applied to the 50 imputed datasets were combined using Rubin's rules (Rubin 1996). Notably, 59% of patients did not have missing data for the selected covariates. Applying multiple imputation allowed for the identification of covariates that affect standard of care effectiveness. Potential predictors for ORR include number of prior lines of therapy, refractory status, thrombocytes, and type of measurable disease. Simulation outcomes further validated the results.

Conclusions This research investigated methodologies for handling missing data in RWD studies and established a clear framework for applying multiple imputation for important covariates within the context of multiple myeloma. The results show that multiple imputation helped to reduce bias and improve validity, reliability, and efficiency of the prediction methods.

20: An Imputation Method for Heterogeneous Studies in Network Meta-Analysis: A Fully Conditional Specification Approach using Distance Metrics

Christos Christogiannis^{1,2}, Dimitris Mavridis²

¹University of Southampton, UK

²University of Ioannina, Greece

Background Multiple Imputation (MI) is a popular method for addressing missing data in Individual Patient Data (IPD) meta-analysis. In an IPD meta-analysis with missing data, the complete case analysis (CCA) is considered a reasonable starting point, and then MI as a sensitivity analysis, and vice versa. Fully Conditional Specification (FSC) is a MI method that addresses missing data by imputing one variable at a time, cycling through iterations of univariate models. In each iteration, the incomplete variable is imputed based both on the complete and previously imputed variables.

Methods Our approach involves estimating the proximity between studies using various distance metrics. By doing so, we identify group studies. Then, we use imputation for each study individually, borrowing information from those studies that exhibit close proximity in terms of distance. Therefore, imputation is informed by neighboring studies, enhancing

its accuracy. We conducted a simulation study to evaluate the properties of the suggested methodology and explore how number of studies, number of patients per study, missing rates, standard deviation of the covariates, heterogeneity, and the correlation of covariates affect the results. After accounting for all the aforementioned factors, we resulted in 216 distinct simulation scenarios. The methods that we compared were CCA, FCS, our proposed approach and the full model as it would be if no missing values were induced in the data. The missing mechanism was set to be MAR.

Results Simulation results were similar between FCS and the proposed method for small percentage of missingness. In scenarios with percentage of missingness of 50% the proposed method outperformed the FCS imputation method in most of the cases. As missingness percentages decreased, our method yielded similar results to FCS, with differences in the third decimal place. More specifically, it had a closer coverage rate (CR) to 95% and was less biased than FCS approach but had a slightly higher root mean square error (RMSE).

Conclusion The proposed method yielded robust results after evaluation. This means that our method may substantially improve estimation when heterogeneous studies are present in IPD meta-analysis.

21: Impact of Lack of Measurement Invariance on Causal Inference in Randomized Controlled-Trials Including Patient-Reported Outcome Measures: a Simulation Study

Corentin Choisy, Yseulys Dubuy, Véronique Sébille

Nantes Université, Université de Tours, INSERM, methodS in Patients-centered outcomes and HEalth ResEarch, SPHERE, 44200 Nantes, France.

Aims Randomized controlled-trials (RCTs) are considered as the gold standard for causal inference. RCTs often include patient-reported outcome measures (PROMs) giving insight into patients' subjective experience regarding e.g., quality of life, fatigue, using questionnaires. PROMs are often treated as any other outcome, e.g. blood pressure, despite having their own specificities. For instance, when measuring fatigue, patients' interpretation of PROM items can differ between groups (Differential Item Functioning, DIF) or change over time (Responses Shift, RS) despite similar fatigue levels. In RCTs, randomization should ensure the absence of DIF at baseline. However, RS may subsequently occur differentially between treatment groups during the study, possibly leading to treatment-related DIF when assessing outcomes post-randomization. While such instances of lack of measurement invariance (MI) may provide a better understanding of patients' experiences, they can also induce measure-

ment bias, if ignored. Our objectives were to measure the impact of lack of MI on causal inference in RCTs and determine how different statistical approaches can handle lack of MI and restore causal inference using a simulation study.

Methods Responses to a PROM were simulated to mimic a two-arm RCT with varying sample size, treatment effect (under H_0 and H_1) and number of items. The number of items affected by DIF and DIF size also varied. Partial credit models (PCM) were used to estimate treatment effect with three strategies: S1: ignoring DIF, S2 and S3: accounting for DIF using two PCM-based iterative procedures, either performing tests on PCM parameters (S2) or an analysis of variance of person-item residuals (S3).

Results When DIF was not simulated, it was not falsely evidenced by S2 and S3. When DIF was simulated and ignored (S1), scenarios under H_0 showed high type-I error rates (up to 74 %), and treatment effect estimations were biased under H_0 and H_1 . Overall, bias increased with the size and the proportion of items affected by DIF.

S2 and S3 helped to reduce DIF impact on bias, type-I error, and restore power in scenarios with a sample size of 600 patients. However, they only provided marginal improvements with smaller sample sizes.

Conclusion This study highlights that causal inference in RCT can be compromised by lack of MI, if ignored or inappropriately recovered. Methods aiming at detecting and accounting for lack of MI can help reduce the risk of biased estimates of treatment effect, particularly when sample size is large.

22: Evaluation of the Psychometric Qualities of Idiographic Patient Reported Outcome Measures (I-PROMs) for Patients Monitoring: PSYCHLOPS Example

Salma Ahmed Ayis¹, Luís Miguel Madeira Faísca², Célia Sales³

¹School of Life Course and Population Sciences; King's College London, United Kingdom

²The University of Algarve, Portugal

³Faculty of Psychology and Education Sciences; University of Porto (FPCEUP)

Introduction/background:

Nomothetic measures are standardised questionnaires that measure patients' self-reported experiences (Patient Reported Outcome Measures (PROMs)). PROMs are brief, acceptable to patients and assessors, and broad enough to capture a breadth of difficulties and

experiences, allowing for population level comparisons. Patients assign scores against norms derived from clinical and non-clinical populations. Change in scores is often used in trials to assess therapeutic effect.

However, nomothetic PROMs are unable to capture unique problems, and circumstances. Patient-Generated Outcome Measures, known as Idiographic PROMs (I-PROMs), allow people to identify their problems, describe these and provide scores to indicate their impact; therefore, allowing the use of appropriate interventions, and the assessment of the efficacy of interventions. The Psychological Outcome Profiles (PSYCHLOPS) is an I-PROM with questions on problems, function, and wellbeing, where patients can describe their problems and their severity scores. WHO have been using PSYCHLOPS for many years as part of their 'Problem Management Plus' intervention.

Nomothetic measures assume that individual questionnaires' items assess one or more underlying construct that can be summarised using latent class-based methods. I-PROMs on the other hand, primarily value the uniqueness of individual experiences, perceptions, and constructions, therefore, using an underlying construct is considered inappropriate in reflecting persons' expressions.

In two studies we examined the theory behind I-PROMs and the potential value of latent class methods in providing an insight into these measures. Factor analysis and Item Response Theory (IRT) were used to understand the properties of PSYCHLOPS, an I-PROM.

Methods Pre- and post-treatment PSYCHLOPS data derived from six clinical samples ($n = 939$) were analysed for validity, reliability and responsiveness; caseness cut-offs and reliable change index were calculated. Exploratory and Confirmatory Factor Analyses were used to determine whether items represented a unidimensional construct; IRT examined items' properties.

Results Estimates for internal consistency, construct validity, and structural validity were satisfactory. Responsiveness was high: Cohen's d , 1.48. Caseness cut-off and reliable clinical change scores were 6.41 and 4.63, respectively. Factor analysis supports items' unemotionality. IRT analysis confirmed that items' scores possess strong properties in assessing the underlying trait measured by PSYCHLOPS.

Conclusion PSYCHLOPS functioned as a measure of a single latent trait, which we describe as 'personal distress'.

There are several challenges for I-PROMs including the robustness of the items to be measured, their measurement model, their reliability and validity, and the meaning of an aggregated I-PROM score. I-PROMs may complement nomothetic measures.

23: Bias in the Estimation of a Psychometric Function when using the PSI-Method under Optimal Conditions – a Simulation Study

Simon Grøntved^{1,2}, Jakob Nebeling Hedegaard¹, Ib Thorsgaard Jensen³, Daniel Skak Mazzhari-Jensen⁴

¹Danish Center for Health Services Research, Department of Clinical Medicine, Aalborg University, Denmark

²Psychiatry, Region North Jutland, Denmark

³Statistics and Mathematical Economics, Department of Mathematical Science, Aalborg University, Denmark

⁴Neural Engineering and Neurophysiology, Department of Health Science and Technology, Aalborg University, Denmark

Background The PSI-method is a Bayesian adaptive method intended to estimate the threshold and slope of a parametrized psychometric function. The method has been used in both research and clinical practice. It was proposed as an improvement over non-adaptive methods due to a potential need for fewer trials before convergence of estimates is achieved. This has resulted in several studies only running 30-40 stimulation trials when estimation was terminated. A similar range of trials was deemed sufficient in the original study presented by the developers of the algorithm.

While concerns about the choice of parametrization for the lapse rate have been raised, the number of trials needed and how this number relates to estimation of thresholds have been under less scrutiny.

Aim We aimed to investigate the potential for bias of the PSI-method's estimates of threshold and slope of the psychometric function, and to investigate whether such bias depended on the number of trials used, the ground-truth threshold, and the slope.

Method We tested the PSI-method (as implemented by the Palamedes toolbox) in a simulation study with 3874 different personality profiles, and 175 simulations per profile. We used a uniform prior for threshold and slope. To restrict potential bias for threshold and slope, we fixed the lapse and guess rates to the ground-truths. We calculated the relative bias in the estimation of threshold along with confidence intervals, and plotted these against the number of trials used, the underlying threshold and slope.

Results We found presence of bias in the estimation of alpha after 50 trials, where the mean relative bias was positive across most person profiles, median 18.1% [IQR: 11.1%, 44.6%], but in the most extreme cases as high as 147.5%. The observed bias was dependent on the ground-truth threshold, and slope. Increasing the number of trials to 150, the relative bias

was considerably reduced to median 4.7% [IQR: 2.6%, 8.9%]. At 1000 trials the relative bias was negligible, median 0.7% [IQR 0.1%, 1.6%], though still mostly positive.

Conclusion Our results indicate the presence of non-negligible bias in threshold estimation when stopping the PSI-method at the typical number of trials used in real-world settings. This bias was found on simulated data under optimal conditions. We thus conclude that the method requires a significantly greater number of trials than typically used. It should be investigated whether these results can be reproduced in a real-world setting.

24: Psychometric Properties Confirmation of the Multiple Sclerosis Autonomy Scale (MSAS) Questionnaire Evaluating Patient Autonomy in Multiple Sclerosis (MS)

Cécile Donzé², Claude Mekies³, Géraud Paillot⁴, Lucie Brechenmacher¹, Alexandre Civet¹, David Pau¹, Delphine Chomette¹, Mikael Cohen⁵, Catherine Mouzawak⁶, Patrick Vermersch⁷

¹Roche SAS, France

²Hôpital saint Philibert, Groupement des Hôpitaux de l'Institut Catholique de Lille Faculté de médecine et de maïeutique de Lille, Lomme, France

³RAMSAY Clinique des Cèdres, Neurologie, CHU Toulouse, Toulouse, France

⁴Association Aventure Hustive, Saint-Malo, France

⁵CRC-SEP Neurologie Pasteur 2, CHU de Nice, Université Côte d'Azur, UMR2CA-URRIS, Nice, France

⁶Structure régionale neuro SEP SYNPASE, Hôpital du Vésinet, Le Vésinet, France

⁷Univ. Lille, INSERM UMR1172 LiNCog, CHU Lille, FHU Precise, Lille, France

Introduction The Multiple Sclerosis Autonomy Scale (MSAS) is a new Patient Reported Outcome (PRO) aiming to evaluate patient autonomy in multiple sclerosis. Our current study's primary objective is to validate the psychometric properties of the MSAS questionnaire.

Methods A longitudinal prospective observational study included MS patients from January 2024 to May 2024 in 33 sites.

The initial MSAS questionnaire contains 10 dimensions in a 36-items short form and has to be completed by patients at inclusion, D15, D30 and up to one year after inclusion (study is still ongoing as of today).

Abstracts of Contributed Posters

Several of the psychometric properties of the MSAS have been evaluated including its construct validity (correlation coefficient between items), Internal consistency (Cronbach's alpha coefficient), unidimensionality (Retrograde Cronbach's alpha curves) and multiidimensionality (multi trait analysis).

This abstract displays the results of the primary objective of the study evaluated at inclusion, with sensitivity analysis carried out at D15 and D30.

Results From the 210 patients included in the study from January 2024 to April 2024, 199 completed the MSAS questionnaire at baseline: 132 (66.3%) with relapsing remitting form of MS (RRMS), 23 (11.5%) with primary progressive (PPMS) and 44 (22.1%) with secondary progressive (SPMS).

Internal consistency: Cronbach's alpha coefficient ranged between 0.59 to 0.96 at inclusion. Removal of one item in dimension with the lowest Cronbach's alpha coefficient led to increase the coefficient in this dimension to 0.67.

Construct validity: Few strong correlation coefficient ($>|0.8|$) between items were observed, and remained between items of the same dimension.

Unidimensionality: Overall, removing impact questions one at a time has no significant impact on Cronbach's alphas. This suggests that the impact questions are highly correlated with each other and are important for the reliability of the scale. The overall Cronbach's alpha coefficient of the questionnaire was 0.845 with 36-items and 0.843 with 35-items.

Multidimensionality: each item was most correlated within its own dimension.

Conclusion Internal consistency was challenged in a dimension and one item had to be removed. The new MSAS-35 items questionnaire is a psychometrically sound measure of autonomy in Multiple Sclerosis.

25: Learning Heterogeneous Treatment Effect from Multiple Randomized Trials to Inform Healthcare Decision-Making: Implications and Estimation Methods

Qingyang Shi¹, Veerle Coupé², Sacha la Bastide-van Gemert³, Talitha Feenstra¹

¹Unit of PharmacoTherapy, -Epidemiology and -Economics, Groningen Research Institute of Pharmacy, University of Groningen, The Netherlands

²Department of Epidemiology and Data Science, Amsterdam University Medical Centers,

Amsterdam, The Netherlands

³Department of Epidemiology, University Medical Center Groningen, University of Groningen, The Netherlands

Evidence synthesis and meta-analysis are crucial for healthcare decision-making, yet it often assumes treatment effects are shared across populations, neglecting heterogeneity by patients' characteristics. This review addresses the critical need to account for heterogeneous treatment effects when synthesizing multiple trials' data to inform decision-making for a specific target population. We present a causal framework for the decision-making process with heterogeneous treatment effects estimated using the data from different sources. We provide an overview of existing methods for estimating these effects from randomized trials, discussing their advantages and limitations in the context of decision-making. The review covers methods utilizing individual patient data (IPD), partly IPD with aggregate data, and exclusively aggregate data. We emphasize the importance of transportability assumptions, such as shared conditional average treatment effect functions and common covariate support, when extrapolating findings from trials to a target population. Furthermore, we discuss value estimation of an optimal treatment rule in the target population, highlighting the necessity of observational data for estimating the baseline function of outcomes. This review aims to guide researchers and practitioners in appropriately applying and interpreting methods for heterogeneous treatment effect estimation that informs healthcare decision-making when using multiple trials' data.

26: Multiple Imputation of Missing Viral Load Measurements in HIV Treatment Trials: a Comparison of Strategies

Tra My Pham¹, Deborah Ford¹, Anna Turkova¹, Man Chan¹, Ralph DeMasi², Yongwei Wang², Jenny O Huang³, Qiming Liao², James R Carpenter¹, Ian R White¹

¹MRC Clinical Trials Unit at UCL, London, United Kingdom

²ViiV Healthcare, North Carolina, US

³GSK, Ontario, Canada

In randomised trials assessing treatments for HIV, a commonly used primary outcome is the proportion of patients achieving or maintaining viral suppression, often defined based on viral load (VL) measurements below a pre-specified threshold, e.g. <400 copies/mL. However, missing data might occur which can impact the analysis of the primary outcome. In addition, in trials of paediatric populations, further complications can arise from measurements being left-censored (i.e. only known to be below a threshold), and obtained from diluted samples

due to insufficient volumes (i.e. the limit of quantification is inflated by the dilution factor). As a result, viral suppression status can become unclear.

Multiple imputation (MI) has been used for handling missing outcome data in trials. However, when a continuous outcome such as VL is dichotomised to define the primary outcome, the imputation model specification requires further consideration. Trial statisticians could impute the missing VL measurements before dichotomising them to determine suppression status, or impute a binary indicator of suppression status directly. Alternatively, MI could be performed such that categories of VL measurements, one of which is the threshold for defining suppression, are imputed.

We aim to explore the performance of these MI strategies for handling missing VL data in a simulation study, in setting with/without left-censoring and dilution. To motivate our simulation study, we use data in ODYSSEY, a trial comparing dolutegravir-based antiretroviral treatment with standard of care in children with HIV.¹ The primary outcome was defined as the proportion of patients with virological or clinical treatment failure by 96 weeks. Here we focus on the virological failure component; for simplicity we define the primary outcome for the simulation study as the first of two consecutive VL measurements of 400 copies/mL. We simulate VL measurements at baseline and multiple follow-up time points to reflect real trial data collection schedules. VL measurements are made missing under both Missing Completely At Random and Missing At Random mechanisms, and missing data are imputed using different MI strategies. Strategies are compared in terms of method failure, bias, standard errors, coverage, power, and type 1 error. The results of this work will provide the basis for recommendations of practical MI strategies that are relevant to statisticians working in HIV treatment trials.

¹Turkova A, White E, Mujuru HA, et al. Dolutegravir as first- or second-line treatment for HIV-1 infection in children. *New England Journal of Medicine* 2021; 385: 2531-2543.

27: A Novel Approach for Assessing Inconsistency in Network Meta-Analysis: Application to Comparative Effectiveness Analysis of Antihypertensive Treatments

Kotaro Sasaki^{1,2}, Hisashi Noma³

¹The Graduate University for Advanced Studies, Japan

²Eisai Co., Ltd., Japan

³The Institute of Statistical Mathematics, Japan

Introduction Network meta-analysis (NMA) is a pivotal methodology for synthesising evidence and comparing the effectiveness of multiple treatments. A key assumption in NMA is consistency, which ensures that direct and indirect evidence are in agreement. When this assumption is violated, inconsistency arises, conceptualized by Higgins et al. [1] as design-by-treatment interactions, where “design” refers to the combination of treatments compared within individual studies. To evaluate inconsistency, various statistical tools have been developed. However, the existing methods based on statistical testing have limitations, including low statistical power and challenges in handling multi-arm studies. Moreover, the testing approaches might not be optimal for inconsistency evaluation, as the primary goal is not to draw definitive conclusions about design-by-treatment interaction but to identify and prioritise specific designs for further investigations into potential sources of bias within the network. To address these challenges, this study proposes a novel approach for evaluating inconsistency using influence diagnostics, focusing on quantifying the impact of individual study designs on the results.

Methods We developed a "leave-one-design-out" (LODO) analysis framework to systematically quantify the influence of individual designs on the overall NMA results. New influence measures were proposed to evaluate these effects comprehensively. To facilitate interpretation, we also introduced the O-value, a summary metric that prioritises designs based on their potential contribution to inconsistency using a parametric bootstrap method. Additionally, a new testing approach was formulated within the LODO framework to identify critical designs requiring further investigation. These methods were applied to an NMA of antihypertensive drugs comprising various study designs.

Results The application of the proposed methods identified key designs contributing to inconsistency in the antihypertensive drug NMA. The influence measures effectively quantified the impact of individual designs. Moreover, the novel testing approach highlighted specific designs warranting further investigation to uncover potential biases. In a sensitivity analysis, excluding trials suspected of causing inconsistency, the rankings of certain treatment effects were reversed.

Conclusion Our proposed method offers an effective framework for evaluating inconsistency in NMA. By enabling the quantitative assessment and prioritisation of individual study designs, it provides deeper insights into the sources of inconsistency and improves the reliability of NMA findings.

References [1] Higgins JP, Jackson D, Barrett JK, Lu G, Ades AE, White IR. Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies. *Res. Synth. Methods.* 2012;3(2):98-110.

28: Investigating (Bio)Statistical Literacy among Health Researchers in a Belgian University Context: A Framework and Study Protocol

Nadia Dardenne, Anh Diep, Anne-Françoise Donneau

Université De Liège Uliege, Belgium

Introduction Even if the literature highlights the importance of developing (bio)statistical literacy (BSL) through curricula and lifelong trainings, few links are made between the BSL and practices of researchers about statistics. However, the causes of statistical misconducts like p-hacking or HARKing are manifold [1] and need to be investigated as a whole with an appropriate BSL framework.

Framework development A BSL framework will be developed and validated based on current (B)SL definitions[2] and the theory of planned behaviour[3] in order to understand the intentional and behavioural demonstrations, i.e. (intentions) to read/perform statistical reports or analyses - when, why, how often and how - through perceived self-efficacy as consumer and producer of statistics, attitudes towards statistics, subjective norm like pressure and practices from colleagues and basic knowledge of statistic. The objectives of the study will be to assess the BSL by investigating associations among the dimensions of the proposed BSL framework. Also, external factors, notably researchers' educational background in statistics, their professional experience and socio-demographic characteristics will be studied in relation to the BSL dimensions.

Methods A cross-sectional study with the population of interest including health scientific and academic staff at Belgian universities will be conducted. The study has been approved by the University Hospital of Liège Ethics committee. The Delphi method will be used to validate some parts of the BSL dimensions while Cronbach α will be computed to assess internal consistency. Further, exploratory and confirmatory factor analysis will be used to validate the factor structure. Structural equation modelling will be employed to analyse the associations between the BSL dimensions, some of which will be treated as latent variable, and to test the effect of the external factors on these dimensions. Statistical analyses will be performed using the statistical package SAS and R with appropriate packages as lavaan.

Conclusion The data collected will enable to establish the links between the BSL dimensions among health researchers at Belgian universities, and to suggest ways forward, particularly in terms of adapting or reinforcing existing BSL curriculum and instructional practices.

1. Hardwicke TE et al. Calibrating the scientific ecosystem through meta-research. *Annu Rev Stat Its Appl.* 2020;7 Volume 7, 2020:11–37.

2. Gal I. Adults' Statistical Literacy: Meanings, Components, Responsibilities. *Int Stat Rev.* 2002;70:1–25.
3. Hein de Vries, Margo Dijkstra PK. Self-efficacy: the third factor besides attitude and subjective norm as a predictor of behavioural intentions. *Health Educ Res.* 1988;3:273–82.

29: Balneotherapy for Peripheral Vascular Diseases: A Systematic Review with a Focus on Peripheral Arterial Disease and Chronic Venous Insufficiency

Mi Mi Ko

Korea Institute of Oriental Medicine, Korea, Republic of (South Korea)

Background Peripheral vascular diseases (PVDs), including peripheral arterial disease (PAD), chronic venous insufficiency (CVI) and coronary artery disease (CAD), significantly impair vascular function and quality of life. Balneotherapy, a non-invasive intervention involving thermal and mineral water therapies, has shown potential benefits in managing these conditions. However, a systematic evaluation of its efficacy remains limited. This systematic review aims to assess the effects of balneotherapy on vascular outcomes, symptom alleviation, and quality of life in patients with PVDs.

Methods A systematic search was conducted in PubMed (Medline), Embase, and the Cochrane Central Register of Controlled Trials (CENTRAL) to identify randomized controlled trials (RCTs) published up to November 2, 2024. The search terms included "Balneotherapy" and "Peripheral vascular diseases," and studies meeting predefined inclusion criteria were selected. Eligible studies focused on patients with PAD and CVI, assessed the effects of balneotherapy, and applied the same adjunct interventions to both treatment and control groups. Data extraction was performed independently by two researchers, and the risk of bias was assessed using the Cochrane Risk of Bias (RoB) tool.

Results A total of 12 RCTs were included in the analysis. For PAD, balneotherapy improved vascular function (e.g., ankle-brachial pressure index, flow-mediated dilation increased walking capacity, and enhanced functional capacity such as leg pain and swelling). In patients with CVI, balneotherapy reduced lower-limb edema, provided pain relief, and improved mobility and quality of life. For CAD, the therapy enhanced endothelial function, reduced vascular inflammation, and improved peripheral perfusion. Adverse events were rare and generally mild, with no severe safety concerns identified. Despite methodological variability, most studies reported favorable effects, particularly in vascular function and symptom management.

Conclusion Balneotherapy appears to be a safe and effective complementary treatment for improving vascular function, walking capacity, and quality of life in patients with PVDs, particularly PAD and CVI. Further large-scale, high-quality trials with long-term follow-up are needed to confirm these findings and optimize treatment protocol.

30: Binomial Sum Variance Inequality Correction of 95% CIs of Percentages in Multicentre Studies Ensures Approximately 95% Coverage with Minimal Width

Paul Talsma¹, Francesco Innocenti²

¹Phastar, United Kingdom

²Maastricht University, The Netherlands

Percentages with corresponding 95% confidence intervals (CI) are often reported for clinical and epidemiologic multicentre studies. Many approaches of centre effect correction exist. These are often complex and/or provide inadequate coverage. A method is presented for constructing the CI which ensures approximately 95% coverage, has minimal width and is not overly complex: the Binomial Sum Variance Inequality (BSVI) correction.

Two studies have been done to investigate coverage and width of intervals using the correction.

In study 1, coverage and width of CIs using the BSVI correction were compared with no correction and with mainstream correction approaches. A simulation study was done where data was generated using binomial distributions with population percentages per centre being the same or differing using pre-specified amounts. CIs were constructed using the Wilson, Agresti-Coull and exact methods, for varying numbers of centres (2-32) and participants per centre (10-160). The ratio of number of participants between centres was systematically changed. Eleven traditional ways of correcting the CI were compared with each other, with no correction and with the BSVI correction. The traditional ways included using the ANOVA, Fleiss-Cuzick, Pearson, Hedeker and GEE methods for estimating the Intra-Cluster Correlation with or without correction for differences in centre size, as well as direct estimation of variances using SAS® (v.9.4) PROC SURVEYFREQ. It was found that intervals constructed with no correction or with traditional methods had coverage which is too high, a finding which could be explained using the BSVI. The BSVI correction was shown to be effective in downwards correcting coverage close to the desired 95% level and reducing interval width.

In study 2, the properties of the BSVI correction for small samples and scarce events were investigated. Data were generated from 2-4 centres with average event percentages: 2, 4, 8,

16, and 32; total N: 6, 12, 24, and 48; mean ratio of centre size: 1, 2, or 3; and differences between centre percentages being none, small, medium, and large using Cohen's effect size. Results show that for N 24 the BSVI correction leads to 95% CIs with adequate coverage (95%) and reduced width compared to no correction. These findings were corroborated with further simulations using the same parameters but N ranging from 14-30 in steps of 2. The BSVI correction is recommended for use for N 24.

Both studies demonstrate that the BSVI correction leads to CIs with adequate coverage and reduced width when compared to other approaches.

31: Sample Size Calculation Methods for Clinical Trials using Co-Primary Count Endpoints

Takuma Ishihara¹, Kouji Yamamoto²

¹Innovative and Clinical Research Promotion Center, Gifu University Hospital

²Department of Biostatistics, School of Medicine, Yokohama City University

Introduction Clinical trials often employ co-primary endpoints to comprehensively evaluate treatment efficacy. In trials where efficacy is established only if all endpoints show significant effects, the Intersection-Union test is commonly applied. While this approach avoids inflation of Type I error rate due to multiple testing, it increases the Type II error rate, necessitating larger sample sizes to maintain adequate statistical power.

However, most trial designs assume independence among endpoints, which may lead to an overestimation of the required sample size. Considering correlations between endpoints can reduce the sample size while maintaining statistical power.

Various sample size determination methods have been developed for co-primary endpoints with different variable types, including continuous, binary, mixed continuous-binary, and time-to-event. Notably, Homma and Yoshida (2023) introduced a method for mixed continuous and count endpoints, but their approach did not address cases where all primary endpoints are count-based.

Objective This study aims to develop a sample size calculation method for clinical trials with co-primary count endpoints.

Methods Co-primary count endpoints often follow different probability distributions, such as Poisson, zero-inflated Poisson (ZIP), and negative binomial distributions. This study

derives analytical expressions to determine the minimum sample size required to achieve statistical significance at a pre-specified nominal significance level while considering endpoint correlations.

Results Simulation studies were conducted under various scenarios to evaluate the impact of endpoint correlation on sample size requirements. The results show that by considering the correlation between endpoints, the required sample size can be greatly reduced, especially when the correlation between endpoints is high.

Conclusion Our proposed methodology provides a practical approach for optimizing sample size determination in clinical trials with co-primary count endpoints. By leveraging endpoint correlations, researchers can design more efficient trials without compromising statistical power. These findings have significant implications for resource allocation and trial feasibility in studies involving co-primary count endpoints.

32: Analysis of Composite Endpoint in Cardiovascular Device Clinical Trials

Hao Jiang, Yonghong Gao

Johnson and Johnson, United States of America

Composite endpoint is often used to assess the safety and the effectiveness of cardiovascular devices to increase study power. For example, MACCE (major adverse cardiac and cerebrovascular events) is commonly used in cardiovascular clinical trials. Time-to-first event analysis, composite event process and Finkelstein Schoenfeld (FS) method are the most used approaches to analyze the composite endpoint to detect the treatment effect of the investigational device.

We investigate the potential power gain or loss of utilizing composite endpoint compared to using only one of the individual component endpoints, under the above mentioned three analysis methods. In addition, we look into the pros and cons of those three methods under different scenarios, including endpoints correlation and censoring mechanism. Simulation studies are conducted to assess the performance of the three methods under different settings. Simulation results are provided which include some thought provoking observations.

33: Bayesian Predictive Monitoring using Two-Dimensional Index for Single-Arm Trial with Bivariate Binary Outcomes

Takuya Yoshimoto^{1,2}, Satoru Shinoda², Kouji Yamamoto², Kouji Tahata³

¹Chugai Pharmaceutical Co., Ltd., Japan

²Yokohama City University

³Tokyo University of Science

Bayesian predictive probabilities are commonly used in phase II clinical trials and can schematically describe the stability of the data in an interim analysis by considering all possible future data. It thus helps researchers make an informed decision about whether a trial should be prematurely terminated or move to phase III trials. Typically, phase II oncology studies follow a single-arm trial, with the primary endpoint being short-term treatment efficacy. Specifically, objective response based on the RECIST guidelines is commonly used as a primary endpoint in terms of the treatment efficacy.

Although the primary endpoint is commonly set as an efficacy outcome, situations may arise in which the safety outcome is equally important as the efficacy outcome. Brutti et al. (2011) presented a Bayesian posterior probability-based approach that imposed a restrictive definition of the overall goodness of the therapy by controlling the number of responders who simultaneously do not experience adverse toxicity. Similarly, Sambucini (2019) proposed a Bayesian decision-making method based on predictive probability, involving both efficacy and safety with binary outcomes.

These strategies are attractive; however, Brutti et al. (2011) and Sambucini (2019) could not capture the difference in a situation where the joint probability of simultaneously being a non-responder to the therapy while experiencing toxicity is substantially different when comparing the results from the historical control and study treatment. Therefore, we propose a novel approach involving a bivariate index vector for summarizing results by considering the joint probability of described above. Also, through the simulation study to evaluate the operating characteristics of design, we show that the proposed method made appropriate interim go/no-go decisions, and made a valuable contribution to the clinical development. For details, see Yoshimoto et al. (2024).

Reference 1: Brutti, P., Gubbiotti, S. and Sambucini, V. (2011). An extension of the single threshold design for monitoring efficacy and safety in phase II clinical trials. *Statistics in Medicine*, 30(14), 1648-1664.

Reference 2: Sambucini, V. (2019). Bayesian predictive monitoring with bivariate binary outcomes in phase II clinical trials. *Computational Statistics & Data Analysis*, 132, 18-30.

Reference 3: Yoshimoto, T., Shinoda, S., Yamamoto, K. and Tahata, K. (2024). Bayesian predictive probability based on a bivariate index vector for single-arm phase II study with binary efficacy and safety endpoints. *Pharmaceutical Statistics*. <http://doi.org/10.1002/pst.2431>

34: Optimising Covariate Allocation at Design Stage using Fisher Information Matrix for Non-Linear Mixed Effects Models in Pharmacometrics

Lucie Fayette^{1,2}, Karl Brendel², France Mentré¹

¹Université Paris Cité, INSERM, IAME, UMR 1137, Paris, France

²Pharmacometrics, Ipsen Innovation, Les Ulis, France

Introduction This work focuses on designing experiments for pharmacometrics studies using Non-Linear Mixed Effects Models (NLMEM) including covariates to describe between-subject variability. Before collecting and modelling new clinical trial data, choosing an appropriate design is crucial. Clinical trial simulations are recommended [1] for power assessment and sample size computation although it is computationally expensive and non-exhaustive. Alternative methods using the Fisher Information Matrix (FIM) [2] have been shown to efficiently optimize sampling times. However, few studies have explored which covariate values provide the most information.

Objectives Assuming a known model with covariate effects and a joint distribution for covariates in the target population from previous clinical studies, we propose to optimise the allocation of covariates among the subjects to be included in the new trial. It aims achieving better overall parameter estimations and therefore increase the power of statistical tests on covariate effects to detect significance, and clinical relevance or non-relevance of relationships.

Methods We suggested dividing the domain of continuous covariates into clinically meaningful intervals and optimised their proportions, along with the proportion of each category for discrete covariates. We developed a fast and deterministic FIM computation method, leveraging Gauss-Legendre quadrature and copula modelling [3]. The optimisation problem was formulated as a convex problem subject to linear constraints, allowing resolution using Projected Gradient Descent algorithm.

We applied this approach for a one-compartment population pharmacokinetic model with IV bolus, linear elimination, random effects on volume (V) and clearance (Cl), and a combined

error (as in [4]). Additive effects of sex and body mass index were included on $\log(V)$, and creatinine clearance on $\log(CI)$. Initial distribution of covariates was imported from NHANES as in [3].

Results Methods were implemented in R using the package PFIM6.1 (<https://cran.r-project.org/web/packages/PFIM/index.html>).

We found that optimal distribution reduces the number of subjects needed (NSN) to get 80% power on relevance or non-relevance of the three covariates. Without constraints, results were intuitive: distribution between extreme intervals only. In a more constrained and realistic setting, optimisation reduced NSN by over 60%.

Conclusion We introduced a novel method to integrate the FIM for NLMEM with covariates to efficiently optimise covariate allocation among patients for future studies. We showed an important reduction in the NSN to achieve desired power in covariate tests.

References

- [1]FDA Guidance for Industry Population Pharmacokinetics. <https://www.fda.gov/media/128793/download>, 2022
- [2]Mentré et al. CPT Pharmacometrics Syst Pharmacol, 2013
- [3]Guo et al. J Pharmacokinet Pharmacodyn, 2024
- [4]Fayette et al. PAGE, 2024

35: Unbiased Estimation for Hierarchical Models in Clinical Trials

Raiann Joanna Hamshaw, Nanxuan Lin

Biostatistics Research Group, Population Health Sciences Institute, Newcastle University, Newcastle upon Tyne, UK

Background In clinical trials, hierarchical models are applied to data where there are dependence between observations occurring within groups, which would violate the independence assumption of some other non-hierarchical estimation methods. These models allow for the incorporation of group analysis as well as individual level analysis. Unbiasedness refers to the identification of an estimator within a class of unbiased estimators that has a uniformly

minimum risk. Researchers often obtain this by minimising the risk for some parameter, and observing whether the result is independent of the parameter.

Methods The modified covariate method proposed by Tian et al. (2014) is a parametric approach to estimating the causal treatment effect (CATE) as well as identifying significant subgroups. This method is shown to be applicable to continuous, binary and survival outcomes. We intend to apply this method to a hierarchical structure, using the benefit of the eliminated nuisance parameter to obtain an unbiased estimate for the overall treatment effect, given an unbiased estimate exists. Simulation studies were undertaken to assess the variance of our treatment effect estimates against that of traditional methods. Sample size estimation calculations for the method were undertaken.

Results Our methods show that the application of the modified covariate method consistently allowed for a treatment effect estimate with smaller variances than that of current methods, even when subgroup sizes were not equal and when the model included small subgroups. The sample sizes needed for this method are lower than that of other frequentist estimation methods, which often obtain more accurate estimates as the sample size and the subgroup sizes increase.

Conclusion This new approach offers advantages over current frequentist and Bayesian methods. The parametric approach to this problem allows for less uncertainty around choosing appropriate parameters than that of the Bayesian methods, as well as having the benefit of having a lower required sample size than that of current frequentist methods. The method obtained smaller variances surrounding the overall treatment effect estimation against both types of methods.

Reference Tian, L., Alizadeh, A. A., Gentles, A. J., Tibshirani, R. A Simple Method for Estimating Interactions Between a Treatment and a Large Number of Covariates. *Journal of the American Statistical Association*, 109:508, 1517-1532. 2014. <https://doi.org/10.1080/01621459.2014.9514>

36: Sample Size Re-Estimation for McNemar's Test in a Prospective Randomized Clinical Trial on Childhood Glaucoma

Markus Schepers¹, Esther Hoffmann², Julia Stingl², Anne Ehrlich³, Claudia Wolf³, Thomas Dietlein⁴, Ingeborg Stalmans⁵, Irene Schmidtmann¹

¹IMBEI, University of Mainz, Germany

²Department of Ophthalmology, University Medical Centre, University of Mainz, Germany

³IZKS Mainz, Germany

⁴Department of Ophthalmology, University Hospital of Cologne, Germany

⁵Department of Ophthalmology, UZ Leuven, Belgium

In clinical trials involving paired binary data, such as those analyzed with McNemar's test, crucial parameters like the correlation between the paired outcomes and the proportion of discordant pairs significantly impact the test's power. However, these parameters are often unknown at the design stage, complicating sample size planning.

We develop sample size re-estimation strategies for McNemar's test, motivated by the PI-RATE study - a prospective, multi-center, observer-blinded clinical trial comparing standard trabeculotomy with micro-catheter assisted trabeculotomy for treating childhood glaucoma. The trial involves centers in Mainz and Cologne (Germany) and Leuven (Belgium).

For fixed effect size, the power of McNemar's test decreases with a higher proportion of discordant pairs and increases with greater correlation between paired observations. However, knowledge about the correlation between paired observations is often limited before the start of a clinical trial. Therefore, adaptive sample size adjustments when interim analyses reveal a certain fraction of discordant pairs is desirable.

We propose practical, generalized recommendations for adaptive designs in studies involving McNemar's test with uncertainty about parameters relevant for power, for re-estimating sample size based on interim estimates of these key parameters. Our recommendations aim at maximizing the conditional power while maintaining the type I error.

37: Bayesian Bivariate Analysis of Phase II Basket Trials Enabling Borrowing of Information

Zhi Cao¹, Pavel Mozgunov¹, Haiyan Zheng²

¹University of Cambridge

²University of Bath

Introduction

Phase II clinical trials focus primarily on establishing early efficacy of a new treatment, while the importance of continued monitoring toxicity cannot be ignored. In the era of precision medicine, basket trials have gained increasing attention, with biomarker-driven technology in various patient sub-populations sharing a common disease feature (e.g., genomic aberration). Thus, the borrowing of information across similar patient (sub-)groups is essential to expedite drug development.

Method

We propose a robust Bayesian hierarchical model that can integrate and analyse clinically rel-

evant differences in toxicity and efficacy, while accounting for possible patient heterogeneity and the correlation between the treatment and toxicity effects. From practical consideration, toxicity responses are treated as binary observations, and the efficacy outcomes are assumed to be normally distributed. Our model can be viewed as a two-dimensional extension of the exchangeable-nonexchangeable (EXNEX^[1]) method: flexible weights are assigned to mixture distributions that imply different borrowing structures concerning toxicity and efficacy, namely, bivariate EX, bivariate NEX, EX in either toxicity or efficacy while NEX in the other.

Results & Conclusion:

Compared with standard Bayesian hierarchical modelling and stand-alone analysis, simulation results of operating characteristics show that our models perform robustly in terms of (the Bayesian analogues of) type I error and power, especially when only toxicity effects are exchangeable (vice versa). The proposed method also has higher power than independently applying the EXNEX method to toxicity and efficacy treatment effects when they are obviously correlated and dissimilar.

Discussion

We give specific model recommendations for various clinical scenarios based on our simulation study of the joint evaluation of treatment effects. Possible future directions to our proposal are the sample size re-estimation and time-to-event extension.

[1] Neuenschwander, Beat et al. "Robust exchangeability designs for early phase clinical trials with multiple strata." *Pharmaceutical statistics* vol. 15,2 (2016): 123-34. doi:10.1002/pst.1730

38: Usefulness of the Blinded Sample Size Re-Estimation for Dose-Response Trials with MCP-Mod

Yuki Fukuyama^{1,2}, Gosuke Homma³, Masahiko Goshio⁴

¹Biostatistics and Data Sciences, Nippon Boehringer-Ingelheim Co., Ltd, Tokyo, Japan

²Graduate School of Comprehensive Human Sciences, University of Tsukuba, Tsukuba, Japan

³Quantitative Sciences, Astellas Pharma Inc, Chuo-ku, Tokyo, Japan

⁴Department of Biostatistics, Institute of Medicine, University of Tsukuba, Tsukuba, Japan

Background / Introduction Sample size calculation requires the assumption of a mean difference and common variance for continuous outcomes. It is often difficult to specify the appropriate value of variance in the planning stage of a clinical trial, and its misspecification results in an unnecessarily large or small sample size. To mitigate such misspecification,

blinded sample size re-estimation (BSSR) has been proposed. Since BSSR uses only accumulated data in a blinded manner for variance estimation, and thus is easy to implement. Several variance estimators have been proposed in BSSR for two-arm trials. Recently, a multiple comparison procedure with modelling techniques (MCP-Mod) has become more common in dose-response trials, as it addresses model uncertainty by specifying a set of candidate dose-response models. Nonetheless, no BSSR in dose-response trials with MCP-Mod has been proposed. We extend variance estimators originally developed in BSSR for two-arm trials, and investigate their usefulness in dose-response trials with MCP-Mod.

Methods For BSSR in dose-response trials with MCP-Mod, we investigate four variance estimators: the unblinded pooled variance estimator, blinded one-sample variance estimator (OS), bias adjusted blinded one-sample variance estimator (bias adjusted OS), and blinded variance estimator using information on randomization block size. We conduct a simulation study to evaluate operating characteristics including the type I error rate and power. Furthermore, to clarify the discrepancy between the actual and nominal power under the final sample size after BSSR, we investigate biases in the point estimates at the end of the trial.

Results BSSR based on the OS can control the type I error rate and ensure the target power even if the true variance differs from the assumed one. On the other hand, BSSR based on the bias adjusted OS and blinded variance estimator using information on randomization block size can control the type I error rate, but cannot always ensure the target power. Furthermore, it is found that the point estimates are biased after the BSSR based on the OS and bias adjusted OS.

Conclusion Although the point estimate is biased after the BSSR based on the OS, it is the only method that satisfies both controlling the type I error rate and ensuring the target power. Therefore, we recommend using the OS-based BSSR to mitigate the misspecification of variance at the trial design stage for dose-response trials. Further investigation on the other endpoints (e.g., binary, count, and time-to-event) may be an avenue for future research.

39: Quantification of Allocation Bias in Clinical Trials under a Response-Adaptive Randomization Procedure for Binary Response Variables

Vanessa Ihl, Ralf-Dieter Hilgers

RWTH Aachen University / Uniklinik Aachen, Germany

Background Randomized clinical trials have as one of their main goals to mitigate bias. The (un-)predictability of assigning a patient to a treatment arm is influenced by different

aspects, one of which is **allocation bias**. It describes the selective allocation of patients by the recruiter influenced by his opinion e.g., on which arm is best or has a higher probability of being allocated. It is based on patient characteristics that influence the expected response, causing bias in the response and therefore affecting the results of a trial.

Response-adaptive randomization (RAR) promises to treat more patients with the more effective treatment compared to classical approaches. Recently, it has been the subject of increased interest and initial use in trials, as it is said to have higher success rates. It is expected that more patients receive the best treatment without compromising the results or requiring more patients for the trial. Specifically for rare diseases, where one expects to include a large proportion of all diseased persons, it is desirable to treat more patients with the better treatment during the phase II/III trials. So far, there is nothing about allocation bias in this area.

Methods We will consider a single-centered two-arm parallel group design with a binary primary endpoint, binomial distributed, in which the doubly adaptive biased coin design is applied for allocating the patients to a treatment arm. Further, we apply a testing strategy that allows to take the adaptive nature of the randomization procedure into account. We implement the procedure for simulations and quantify the allocation bias with special focus on rare diseases. Different assumptions for the allocation bias are investigated including strict biasing policies and higher values for the effect of the bias.

Results Our first results indicate that even if the allocation bias can be very strong, some RAR procedures are hardly influenced. Generally, the responses in a simulation study seem to be weakly affected by allocation bias. For specific strategies however, it is important to model possible bias effects. Further simulations are still in process, the upcoming results are expected to strengthen this hypothesis.

Conclusion RAR trials can successfully reduce concerns about allocation bias for certain procedures. In some cases, it is useful to be able to include modelling the bias in the trial analysis if it cannot be addressed initially in the design of the trial.

40: Assessment of Assay Sensitivity in Non-Inferiority Trials Using Aggregate Data from a Historical Trial: A Population Adjustment Approach

Eisuke Hida¹, Satomi Okamura², Tomoharu Sato³

¹The University of Osaka, Japan

²The University of Osaka Hospital

³Hiroshima City University

Background In non-inferiority (NI) trials lacking assay sensitivity, an ineffective treatment may be found non-inferior, potentially leading to an erroneous efficacy assessment. Therefore, a 3-arm NI trial with placebo, test treatment and control treatment is considered the gold standard and is recommended by several guidelines, such as ICH-E10. However, due to ethical and feasibility concerns regarding the inclusion of a placebo, the practical implementation of 3-arm NI trials remains limited. As a result, a useful method for evaluating assay sensitivity in 2-arm NI trials is needed.

Objective We propose a practical method for confirming assay sensitivity in 2-arm NI trials. This method evaluates the assay sensitivity of the NI trial after adjusting for the distribution of covariates using a population adjustment method, applied to the summary statistics of historical trial data and the individual patient data (IPD) of the NI trial.

Method To assess assay sensitivity, it is necessary to demonstrate that the acceptable minimum effective value of the test treatment in a 2-arm NI trial is superior to the placebo (Hida & Tango, 2018). Since a placebo is not included in 2-arm NI trials, historical trial results must be used as external information. To evaluate assay sensitivity in NI, an adjustment method is required to align the patient characteristics from a historical trial to the distribution of IPD in the NI trial. In other words, this approach is the reverse of Matching Adjusted Indirect Comparison (MAIC) or Simulated Treatment Comparison (STC). This proposed method evaluates assay sensitivity of the NI trial by estimating the average treatment effect of a historical trial in the population of the NI trial (IPD) through a combination of MAIC and inverse probability weighting (IPTW). We investigated the performance and practicality of the proposed methods through simulations based on several scenarios using clinical trial data.

Results and Conclusions Although the proposed method relies on external information, which may result in a lower level of evidence compared to the gold standard design, this method suggests that it is useful for evaluating assay sensitivity in NI trials and supporting decision-making.

41: Exploring Methods for Borrowing Evidence Across Baskets or Subgroups in a Clinical Trial: a Simulation Study

Wenyue Li, Becky Turner, Duncan Gilbert, Ian White

University College London (UCL), United Kingdom

Introduction Basket trials are designed to study a single targeted therapy in the context of multiple diseases or disease types sharing common molecular alterations. To draw adequate inference about small baskets, approaches for borrowing evidence become crucial. We aimed to quantify the benefits of information borrowing and to compare the performance of various methods for a proposed phase III basket trial studying a novel immunotherapy for patients with mucosal squamous cell cancers in two common and four rare cancer sites.

Methods We simulated six scenarios with different patterns of variation in true treatment effects of a time-to-event outcome across sites. Scenarios 1, 2 and 3 assumed high, low and moderate variation, while Scenarios 4 and 5 assumed contradictory data for the common sites with high or low variation among the remaining sites, respectively. Scenario 6 assumed similar estimates for the common sites only. We estimated a two-stage random-effects meta-analysis model using restricted maximum likelihood or Bayesian estimation while incorporating different priors for the between-site variance. We also implemented an empirical Bayes method and one-stage Bayesian hierarchical approaches using a Bayesian hierarchical model (BHM) and an exchangeability-nonexchangeability (ExNex) model. We conducted 1000 simulations to compare the performance of all methods to a standalone analysis.

Results The standalone method performed the worst in precision, mean squared error and power despite its robustness for bias. On the other hand, the Bayesian meta-analysis method with a strongly informative prior was the most precise while producing very large biases under most scenarios, except Scenario 2 where the empirical Bayes method appeared to be the most precise. However, a substantial undercoverage was found for the empirical Bayes method under Scenario 2 and for the Bayesian meta-analysis method with a strongly informative prior under Scenarios 1, 4, 5 and 6. The ExNex model resulted in fairly low biases under most scenarios, whereas the BHM achieved considerably higher precision and power than the former for the rare sites under Scenario 2.

Conclusions Our work demonstrated precision and power gains from using proposed information-borrowing methods rather than a standalone analysis. We also demonstrated sensitivity of the results to the choice of prior for the between-site heterogeneity. To provide further guidance for practice, we recommended using a vague prior for the Bayesian meta-analysis method when treatment effect heterogeneity is likely to be limited. Moreover, we recommended using the ExNex model when contradictory true treatment effects are likely to exist.

42: Comparing the ED50 Between Treatment Groups Using Sequential Allocation Trials.

Teresa Engelbrecht, Alexandra Graf

Medical University Vienna, Austria

Determining the median effective dose (ED50) is a fundamental objective in the field of anaesthesia research, both in human and animal studies. Sequential allocation methods, such as the Up-and-Down Method (UDM) and the Continual Reassessment Method (CRM), offer an efficient method of dose allocation based on the responses of previous subjects, thus reducing the required sample size compared to traditional study designs with fixed sample sizes [1].

Motivated by previous studies [2,3], we aim to evaluate methods for comparing ED50 across different treatment groups. While sequential allocation methods such as the Up-and-Down Method (UDM) and the Continual Reassessment Method (CRM) are well described for estimating the ED50, only a limited amount of literature is available for the comparison of the ED50 between several treatment groups. To evaluate the advantages and limitations of sequential allocation methods in comparison to traditional fixed-sample designs, we conducted simulation studies across various scenarios. Our analysis assessed the power and type-1-error of UDM and CRM, as well as logistic regression with a fixed sample size, to determine their respective strengths and weaknesses in estimating and comparing ED50 values across treatment groups.

[1] Görge M, Zhou G, Brant R, Ansermino JM. Sequential allocation trial design in anaesthesia: an introduction to methods, modeling, and clinical applications. *Paediatr Anaesth*. 2017;27(3):240-247. doi:10.1111/pan.13088

[2] Müller J, Plöchl W, Mühlbacher P, Graf A, Stimpfl T, Hamp T. The Effect of Pregabalin on the Minimum Alveolar Concentration of Sevoflurane: A Randomized, Placebo-Controlled, Double-Blind Clinical Trial. *Front Med (Lausanne)*. 2022;9:883181. Published 2022 May 3. doi:10.3389/fmed.2022.883181

[3] Müller J, Plöchl W, Mühlbacher P, Graf A, Kramer AM, Podesser BK, Stimpfl T, Hamp T. Ethanol reduces the minimum alveolar concentration of sevoflurane in rats. *Sci Rep*. 2022;12(1):280. Published 2022 Jan 7. doi:10.1038/s41598-021-04364-8

43: A Pre-Study Look into Post-Study Knowledge: Communicating the Use(Fulness) of Pre-Posteriors in Early Development Design Discussions

Monika Jelizarow

UCB Biosciences GmbH, Germany

When designing a clinical study we make assumptions on our drug's true treatment effect, for the endpoint of interest. These assumptions are based on existing data and/or expert belief, that is, they are based on *some* form of evidence synthesis. In the Bayesian framework, this evidence synthesis will result in a design prior distribution representing our *current* knowledge about the true treatment effect. A pre-posterior can be interpreted as a conditional posterior distribution representing the *updated* knowledge about the true treatment effect at the end of our future study given only that we know that a certain study outcome (i.e. success or failure) has been met (Walley et al., 2015; Grieve, 2024). Thus, pre-posteriors enable a look into future updated evidence (this is the 'post' part) before running the future study (this is the 'pre' part). This opens the door to help answer many questions statisticians are often asked by their clinical colleagues in proof-of-concept settings, e.g. '*If the study will be successful, what will this make us learn about the true treatment effect?*' or '*Does running the study de-risk our program? How would it compare to running the study with more (or fewer) patients?*' Shaped by experiences gained in our organisation, the goal of this contribution is to propose a question-led and visualisation-informed workflow for how to effectively communicate the (use)fulness of this quantitative tool in discussions with stakeholders. The importance of early contextualisation will be emphasised, and supported by illustrating the relationship between, for example, the pre-posterior of success and the unconditional probability of success (PoS), also known as assurance.

44: Estimation and Testing Methods for Delayed-Start Design as an Alternative to Single-Arm Trials in Small Clinical Trials

Tomoharu Sato^{1,2}, Eisuke Hida²

¹Hiroshima City University, Japan

²The University of Osaka, Japan

Introduction and Objective(s):

Traditional randomised controlled trial designs are difficult to implement in small populations, such as in the rare disease and paediatric disease areas. Various methodological and statistical considerations have been reported for such small clinical trials [1, 2]. Due to feasibility, many single-arm trials of test drugs alone are still being conducted, allowing the evaluation of within-patient comparisons. In single-arm trials, the efficacy of a test drug is assessed based on a pre-specified threshold. However, it is well known that even if the treatment effect is better than the threshold in a well-controlled single-arm trial, the estimate of the treatment effect is subject to bias. Therefore, simple estimates from single-arm trials may make it difficult to draw valid conclusions about efficacy. In such situations, it is also desirable to be

able to estimate the true effect size of the test drug without the influence of bias. In this study, we propose a delayed-start design as an alternative to single-arm trials and a method for estimating and testing treatment effects.

Method(s) and Results:

We propose a method for estimating and testing treatment effects using a delayed start design. In a delayed start design, a randomised controlled trial is conducted in the first period and a single-arm trial in the second period, allowing to estimate the treatment effect of the trial and the difference between the two treatment effects. Various factors, such as disease and treatment characteristics, determine the 'estimand' and alter the modelling, but we have given model-specific estimation and testing methods and interpretations. We show that, with appropriate use of delayed start designs, it is possible to estimate the difference between two treatment effects, in addition to assessing efficacy by comparison with a pre-specified threshold, as in single-arm trials. Numerical study is also used to assess their performance and give model-specific interpretations.

Conclusions Delayed start designs with appropriate modelling for the primary endpoints may be more effective than single-arm trials for pragmatic small clinical trials in rare and paediatric disease areas.

Keywords small clinical trials, delayed-start design

References [1] IOM. Small clinical trials. issues and challenges (2001).
[2] CHMP. Guideline on clinical trials in small populations (2006).

45: Dealing with Missing Values in Adaptive N-of-1 Trials

Juliana Schneider¹, Maliha Raihan Pranti², Stefan Konigorski^{1,3,4}

¹Hasso-Plattner-Institute, Germany

²University of Potsdam, Germany

³Hasso Plattner Institute for Digital Health at Mount Sinai

⁴Icahn School of Medicine at Mount Sinai

N-of-1 trials are multi-crossover trials in single participants, designed to estimate individual treatment effects. Participants alternate between phases of intervention and one or more alternatives in

trials that often have only few data points. In response-adaptive designs of N-of-1 trials, trial length and burden due to ineffective treatment can be reduced by allocating treatments adaptively based on interim analyses. Bayesian approaches are directly applicable by updating posterior beliefs about effectiveness probabilities. Furthermore, they allow inference for both individual and aggregated effects. Missing values occur, for instance, due to commonly reported wavering adherence to the treatment schedule and other personal or external factors. This may happen randomly throughout the trial (Missing Completely At Random) or dependent on other factors such as severity of symptoms addressed in the trial or time. Missing values require adjusting the adaptive allocation mechanism appropriately, but the best approaches for short adaptive N-of-1 trials are not known. In fact, careful imputation of such missing values is crucial, since sequential treatment allocation depends on past outcome values. Here, we investigate the performance of different imputation methods for missing values in simulated adaptive N-of-1 trials. The imputation approaches use information either from only the respective individual or from all participants, and the adaptive N-of-1 trials are set up in a Bayesian-bandit design using Thompson Sampling. We evaluate the different imputation approaches in a simulation study of 1000 synthetic adaptive n-of-1 trials, comparing two alternate treatments and their association with a normally distributed outcome. We compare posterior descriptive and inference metrics for adaptive trajectories with and without missing values. More precisely, we juxtapose the posterior means and variances of the fully observed and partly observed trial sequences against each other and the underlying true distribution, as well as study the Kullback-Leibler divergences among them. This serves to investigate the impact of data missingness and different imputation methods on bias and efficiency in treatment effect difference estimation.

Preliminary results indicate that the optimal imputation method in a given situation depends on whether analysis is intended on the aggregated or individual level. Moreover, the amount of missingness within and between trial participants impacts imputation results. Lastly, time-dependent associations between measurements and missingness may alter the success of various imputation methods. Future research may include such time-dependencies both in the simulated data as well as in suitable imputation methods.

46: Adaptive Clinical Trial Design with Delayed Treatment Effects using Elicited Prior Distributions

James Salsbury¹, Jeremy Oakley¹, Steven Julious¹, Lisa Hampson²

¹University of Sheffield, United Kingdom

²Advanced Methodology and Data Science, Novartis Pharma AG, Switzerland

Randomized clinical trials (RCTs) are essential for evaluating new treatments, but modern therapies such as immunotherapies present challenges, as delayed treatment effects often occur. These delayed effects complicate trial design by leading to premature futility decisions or inefficient trials with excessive sample sizes and extended durations. Additionally, the proportional hazards assumption, commonly used in survival analysis, may be violated in the presence of time-varying treatment effects.

Adaptive trial designs provide a flexible alternative, allowing modifications based on accumulating data. However, in the context of delayed treatment effects, incorporating prior knowledge about uncertain parameters, such as delay duration and effect magnitude, can significantly enhance trial efficiency. Eliciting prior distributions for these parameters provides a structured approach to account for uncertainty, helping guide trial decisions and improve design robustness.

We present a framework for adaptive clinical trials that explicitly incorporates elicited priors to account for delayed treatment effects. We propose adaptive strategies such as dynamic interim analysis, and efficacy/futility stopping rules, which can be informed by prior distributions. Simulations compare the performance of adaptive designs to traditional fixed designs, demonstrating the benefits of using priors to improve trial efficiency and decision-making.

Abstracts of Contributed Posters

Our methods aim to reduce inefficiencies and support real-time decision-making, ultimately advancing the evaluation of new therapies.

Monday Posters at ETH

Monday, 2025-08-25 10:45 - 11:30, ETH, UG hall

1: Retracted

2: Cardio-Metabolic Traits and Its Socioeconomic Differentials among School Children Including MONW Phenotypes in India: A Baseline Characteristics of LEAP-C Cohort

Kalaivani Mani¹, Chitralok Hemraj¹, Varhlunchhungi Varhlunchhungi¹, Lakshmy Ramakrishnan¹, Sumit Malhotra¹, Sanjeev Kumar Gupta¹, Raman Kumar Marwaha², Ransi Ann Abraham¹, Monika Arora³, Tina Rawal⁴, Maroof Ahmad Khan¹, Aditi Sinha¹, Nikhil Tandon¹

¹All India Institute of Medical Sciences, Delhi, India

²International Life sciences Institute, Delhi, India

³Public Health Foundation of India, Delhi, India

⁴HRIDAY, Delhi, India

Background Cardio-metabolic risks emerge in early life and are transmitted into adult life. Further, these risks may have aggravated due to worsening food security and diet quality during the pandemic. We aimed to assess the prevalence of cardiometabolic traits including the metabolically obese normal weight phenotype and socioeconomic differentials in children and adolescents aged 6-19 years in India.

Methods A baseline assessment was conducted between August 17, 2022, and December 20, 2022, as part of a school-based cohort study that aimed at longitudinally evaluating the anthropometric and metabolic parameters among urban children and adolescents aged 6-19 years from three public schools and two private schools in India. Private and public schools were considered a proxy for higher and lower socioeconomic status respectively. Blood pressure and blood samples in a fasting state were obtained only from adolescents. The prevalence along with its 95% confidence interval using Clopper exact method and adjusted prevalence ratios was calculated using random-effects logistic regression models.

Results Among the 3,888 students (aged 6–19 years) recruited, 1,985 were from public schools and 1,903 from private schools. The prevalence of underweight was 4.95% (95% CI 1 · 25-12 · 72), significantly higher in public schools ($p<0.0001$), while general obesity (13.41% (95% CI 2 · 98-33 · 87)) and central obesity (9.15% (95% CI 1 · 40-27 · 44)) were significantly higher in private schools (adjusted PR = 4.42 and 8.31, respectively). Hypertension prevalence (7.37% (95% CI 6 · 44-8 · 38)) was similar across schools, but impaired fasting glucose (adjusted PR = 2.37) and metabolic syndrome (adjusted PR = 3.51) were more common in private schools. Among 2,160 adolescents, 67.73% had a normal BMI, with a 42.86% (95% CI 30 · 79-55 · 59) prevalence of the metabolically obese normal weight (MONW) phenotype, higher in public (46.39%) than private (35.33%) schools ($p=0.0742$). Low HDL-C was the most common MONW abnormality (41.74%), significantly more prevalent in public schools (62.12% vs. 52.73%, $p=0.0393$).

Conclusion Effective implementation of food security measures and targeted initiatives will be crucial to mitigate the socio-economic and gender disparities associated with the growing burden of cardiometabolic traits. Metabolic obesity among phenotypically normal or underweight adolescents should not be overlooked but intervened early through novel screening criteria to prevent future cardiovascular burden. These findings also have implications for low-income and middle-income countries like India undergoing nutritional transition where socioeconomic status strongly influences cardio-metabolic traits.

3: External Validation of SMART2 Model for Recurrent Cardiovascular Risk

Jasper Wilhelmus Adrianus van Egeraat, Nan van Geloven, Hendrikus van Os

LUMC, Netherlands, The

Background Assessing performance of prediction models in external data is important before use in medical practice. In real medical data sets, this may be challenged by several data complexities, including censoring, competing events and missing data. For example, when using routine electronic health records, the missing at random (MAR) property required for multiple imputation is often violated, possibly leading to inaccurate performance metrics.

This work illustrates how the combined challenges of censoring, competing events and missing data were addressed when evaluating the predictive performance of the SMART2 prediction model. The SMART2 prediction model can identify individuals at high risk of recurrent atherosclerotic cardiovascular diseases.

Methods Electronic health records from the Extramural LUMC Academic Network were used

to derive routine clinical data from patients registered between January 2010 and December 2021 in the greater Leiden-The Hague region of the Netherlands. Individuals were included if they had been hospitalized for cardiovascular disease. The outcome was the first recurrent occurrence of a composite of non-fatal myocardial infarction, non-fatal stroke, and vascular death within 10 years.

Calibration plots and observed/expected (OE) ratios were determined. Censoring and competing events were incorporated in the observed outcome proportion with the Aalen-Johansen estimator. Discrimination was determined between subjects who developed the primary event before 10 years and those who did not experience any event by 10 years, applying inverse probability of censoring weights.

Missing variables were handled using multiple imputation with chained equations. Longitudinal measurements were used to improve imputation of the measurements used at the prediction moment. To account for possible missingness not at random, a sensitivity analysis was performed by delta-scaling the imputed values after each iteration, mimicking various degrees of missingness not at random.

Results Out of the 15,561 included patients, 2,257 patients suffered a recurrent cardiovascular event and 2,098 had a competing event. The median follow-up time was 6.07 years. The AUC_t was 0.62 (95%CI: 0.60–0.64) and the OE ratio was 0.97 (95%CI: 0.93–1.02).

Discrimination was robust under various delta-scaling parameters. Assuming unobserved predictors were overestimated by the imputation model, scaling imputed values downward by 10% every iteration, resulted in an AUC_t of 0.62 (95%CI: 0.60–0.64). The OE ratio changed to 1.01 (95%CI: 0.96–1.05).

Conclusions In this real-world analysis challenged by censoring, competing events and missing data, we showed the feasibility of testing robustness of predictive performance assessment under varying degrees of missingness not at random.

4: Multi-Disease Risk Models to Target Concomitant Diseases and their Interactions: Insights on Cardio-Renal-Metabolic Syndrome in England

Stelios Boultsakis Logothetis¹, Niels Peek², Angela Wood¹

¹British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, Victor Phillip Dahdaleh Heart and Lung Research Institute, University of Cambridge, United Kingdom

²THIS Institute (The Healthcare Improvement Studies Institute), University of Cambridge, United Kingdom

Introduction Clinical risk prediction models are used to identify patients at high risk of disease onset. However, most existing approaches only focus on a single disease, ignoring clusters of conditions with shared pathophysiology and common treatments. Accounting for these relationships could support better disease prevention and health outcomes.

This study develops multi-disease models to jointly predict cardiovascular disease, chronic kidney disease, and metabolic disorders like diabetes. These conditions, collectively termed cardio-renal-metabolic syndrome, share risk factors and intervention effects and are significant contributors to premature mortality. We aim to extract insights about disease progression in the English population and lay the foundations for future individualised multi-disease prediction models.

Methods We modelled disease progression as a state transition process, fitting a multi-state model to predict 5-year incident cardiovascular disease (CVD) and chronic kidney disease (CKD), with diabetes as a risk factor and death as a competing risk. State transition intensities were jointly estimated using Cox proportional hazards sub-models.

We extracted a novel dataset of electronic health records spanning the entire adult population of England from NHS databases, including diagnoses, laboratory measurements, and treatments. Missing data were multiply imputed, and we ensured congeniality with the multi-state model by including non-parametric state probabilities in the imputation. To support computational feasibility, we discretised and coarsened the time scale and restricted to a curated set of well-established risk predictors.

Results We identified 394,555 cases of concomitant CVD and CKD among the 48.65 million eligible adults. The incidence of CKD following a CVD diagnosis was approximately twice that of CVD following a CKD diagnosis (24.73 vs. 12.85 per 1000 person-years). The Cox models achieved an average concordance index of 0.882 across imputations. Nearly all predictors were significantly associated with every state transition. The strongest predictor was smoking, with hazard ratios ranging from 2.14-2.69.

Conclusion We demonstrated how cardio-renal-metabolic syndrome can be jointly modelled at a national scale. Next, we will experimentally evaluate this model's individual-level predictions and develop more granular multi-state models that include additional clinically relevant intermediate states. The optimisations required for model fitting suggest that classical approaches are reaching their computational limits. Future work will explore machine learning methods to better leverage whole-population electronic health records and their wide range of risk predictors.

5: Machine Learning Methods for Analyzing Longitudinal Health Data Streams: A Comparative Study

Inês Sousa

Universidade do Minho, Portugal

Chronic kidney disease (CKD) is characterized by kidney damage or an estimated glomerular filtration rate (eGFR) of less than 60 ml/min per 1.73 square meters for three months or more. The performance of six tree-based machine learning models - Decision Trees, Random Forests, Bagging, Boosting, Very Fast Decision Tree (VFDT), and Concept-adapting Very Fast Decision Tree (CVFDT)- are evaluated on longitudinal health data. Longitudinal data, where individuals are measured repeatedly over time, provide an opportunity to predict future trajectories using dynamic predictions that incorporate the entire historical dataset. These predictions are essential for real-time decision-making processes in healthcare. The dataset comprised 406 kidney transplant patients, spanning from January 21, 1983, to August 16, 2000. It captures 120 time points over the first 119 days post-transplant, including baseline glomerular filtration rates (GFR), along with three static variables: weight, age, and gender. Data preprocessing involved robust imputation techniques to handle missing data, ensuring consistency and trend accuracy. The models were trained to predict health outcomes starting from the eight-day post-transplant, progressively incorporating daily values to predict subsequent days up to day 119. Model performance was evaluated using mean squared error (MSE) and mean absolute error (MAE) through data partitioning and cross-validation techniques.

6: Evaluating the Fairness of a Clinical Prediction Model for Outcomes Following Psychological Treatment in the UK's National Health Service

Nour Kanso¹, Thalia C. Eley¹, Ewan Carr²

¹Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK

²Department of Biostatistics and Health Informatics, Institute of Psychology, Psychiatry and Neuroscience, King's College London, London, UK

Background

Depression and anxiety are common psychiatric conditions that significantly affect individuals' well-being. The UK NHS Talking Therapies programme delivers evidence-based psycholog-

ical treatments to over a million patients annually, but outcomes are heterogeneous; only half achieve clinical definitions of recovery. Stratified care involves predicting outcomes using patient characteristics to identify individuals who may need adapted or alternative treatments. However, the fairness, accuracy, and generalisability of such prediction models across sociodemographic subgroups remain underexplored. This study evaluates the stability and performance of an existing clinical prediction model for outcomes following treatment across gender, employment status, ethnicity, age, and sexuality.

Methods

We evaluated an existing clinical prediction model across sociodemographic subgroups to assess prediction stability and performance variations. Outcomes included reliable improvement in depression (PHQ-9) and anxiety (GAD-7), defined as a change from baseline to the end of treatment exceeding the measurement error of the scale (6 points for depression; 4 for anxiety). Predictors included age, gender, ethnicity, religion, language proficiency, employment, sexuality, long-term condition, disability, medication, prior referrals, diagnosis, and symptom severity. Stability was assessed using bootstrapping (200 iterations), where the model was repeatedly trained on resamples of the dataset and tested within sociodemographic subgroups. Sample size calculations suggested a minimum of 1,788 participants per subgroup, assuming 50% prevalence and a c-statistic of 0.7. Performance was evaluated across subgroups based on calibration and prediction instability.

Results

The analytical sample ($n = 30,999$) was predominantly female (73%) with a median age of 34, and had an ethnic composition including 57% White, and 22% Black, Black British, Caribbean, or African. In the full sample, the model demonstrated good discrimination (depression AUC: 0.76, anxiety: 0.75) and calibration (intercept/slope: -0.00/0.99 (depression), -0.02/1.03 (anxiety)). We observed differences in performance and stability across subgroups. Model calibration and stability were higher for women, whereas the model tended to underestimate outcome probabilities for men. The model also underestimated the probability of reliable improvement for unemployed and retired individuals, especially at the extremes of the probability range. Our full results will present differences by ethnicity, age, and sexuality.

Conclusion

No study to date has explored the fairness of clinical prediction models for psychological therapy in the UK NHS. Our study addresses major gaps in understanding predictive performance across sociodemographic subgroups within UK NHS Talking Therapies. By evaluating fairness, accuracy, and stability, findings will inform model refinements, supporting equitable and reliable treatment recommendations.

7: Multiple Imputation vs. Machine Learning for Handling Missing Data in Prediction Modelling: Which Best Balances Stability, Performance, and Computational Efficiency?

Pakpoom Wongyikul¹, Phichayut Phinyo¹, Noraworn jirattikanwong¹, Natthanaphop Isaradech², Wachiranun Sirikul², Arintaya Phrommintikul³

¹Department of Biomedical Informatics and Clinical Epidemiology (BioCE), Faculty of Medicine, Chiang Mai University, Chiang Mai, Thailand

²Department of Community Medicine, Faculty of Medicine, Chiang Mai University, Chiang Mai, Thailand

³Division of Cardiology, Department of Internal Medicine, Faculty of Medicine, Chiang Mai University, Chiang Mai, Thailand

Background Missing data is a common challenge in clinical prediction modelling. Multiple Imputation with chained equation (MICE) remains the main approach but is computationally intensive and adds complexity. Recent evidence suggests that simpler machine learning-based methods may perform just as well. This study compares MICE and machine learning-based approaches for handling missing data in terms of prediction stability, performance, and computational time to identify the most balanced approach.

Methods A real-world dataset of 8,245 patients, previously used to develop a clinical prediction model for major adverse cardiovascular events, was utilised. We then generated nine datasets to represent different missing data scenarios, varying by missing variable type (categorical, continuous, or mixed) and missing proportion (20%, 40%, or 60%). All missing data were assumed to be missing at random (MAR). Four methods to handle missing data were evaluated: (1) MICE, (2) random forest (RF), (3) k-nearest neighbor (kNN), and (4) complete case analysis (CCA). Performance and stability were evaluated using the bootstrap internal validation procedure according to Riley and Collins. Model performance was assessed with optimism-corrected area under the curve (AUC) and calibration slopes, while stability was measured using mean absolute prediction error (MAPE). Bootstrapping time was also recorded and compared.

Results With 20% missing data, RF, MICE, and kNN showed comparable AUC and MAPE, though kNN exhibited poorer calibration. As missing data increased, all methods except CCA maintained similar AUC, but prediction stability declined, particularly for mixed variable types. Across all scenarios, MICE performed best overall, followed by RF. While kNN produced stable predictions with high AUC, significant miscalibration persisted in most cases, except when 20%–40% of continuous data was missing. In terms of computational efficiency, MICE was the most intensive, taking two to three times longer than RF and kNN.

Conclusions Provided the development sample size is sufficiently large, RF is preferred for its balance of predictive performance, stability, and computational efficiency. If computational time is not a constraint (e.g., with access to high-performance computing), MICE is recommended, followed by RF. Otherwise, kNN may be a suitable alternative when missing data are continuous and below 40%. Finally, CCA should be avoided in all cases.

8: Estimating Rates of Dental Treatment and Unmet Dental Need in a Spatially Explicit Model in Children in England, 2016-2018

Beatrice Catherine Downing

University of Bristol

Registries and the extensive collection of linked data has led to extraordinary advances in our understanding of disease dynamics and optimal resource allocation. However, this requires accompanying investment, collaboration and continuity at multiple levels over many years. In an imperfect world, unlinked data and aggregate counts are more readily available. With proper communication of the uncertainty, aggregate unlinked data from different sources can be used to estimate and validate the prevalence of disease and the scale of unmet clinical need. Here we use publicly available data on the number of dental procedures in children and a signifier of unmet need - the number of hospitalisations for tooth extraction - to estimate relative rates of dental ill-health and to identify areas in England with relatively high levels of unmet need given the level of background deprivation. We used Bayesian hierarchical spatial models to allow for spatial correlation between neighbouring areas, bringing together dental procedures at fine scales and hospital extractions at coarse scales. We demonstrate the power of modelling spatial relationships in systems where both service provision and wider determinants of health show spatial structuring.

9: Statistical Approach to Assess the Impact of Hospital Settings on Optimal Staffing Levels

Diana Trutschel¹, Maryam Ahmadi Shad¹, Michael Ketzer¹, Jack Kuipers², Giuseppe Moffa³, Michael Simon¹

¹Department of Public Health, Institute of Nursing Science, University Basel, Switzerland

²Department of Biosystems Science and Engineering, ETH Zürich, Switzerland

³Department of Mathematics and Computer Science, University of Basel, Switzerland

Background Optimal hospital staffing, often measured by the patient-to-nurse ratio (PNR), is critical to healthcare quality and patient outcomes. Variations in PNR are driven by factors originating from both the patient and nursing side of the ratio. Understanding the extent to which these factors influence PNR is essential for designing effective strategies to achieve and sustain optimal staffing levels. Identifying the relative contributions of these influences can guide decision-making by highlighting the potential impact of adjustments within the healthcare setting.

Methods The distribution of PNR was derived through theoretical modeling, incorporating the relationship between the number of patients and available nursing staff, and approximate real-world PNRs. Simulations were conducted to explore the impact of key variables such as planned staffing schemes and staff absence rates (80, 85, 90, 95%) representing various healthcare settings presented by unit size (20, 30, 40 beds) and occupancy rate (70, 80, 90%). These simulations estimated the proportion of days with overstaffing and understaffing by calculating the area under the PNR distribution curve for values outside a predefined optimal PNR range. This approach enabled the quantification of deviations from optimal staffing levels across diverse scenarios, providing insights into the sensitivity of PNR to changes in system parameters.

Results The simulation results indicate that most common staffing configurations exhibit a high risk of understaffing compared to standard PNR schemes. In a 20-bed unit with a nurse absence rate of 80%, more than 50% of hospital days show overstaffing for PNR values of 6 or higher, whereas more than 50% show understaffing for PNR values of 4 or lower. The findings further demonstrate that variations in staffing plans and nurse absence rates affect the proportion of over- and understaffed days. Smaller units (e.g., 20 beds) are more prone to overstaffing, with nurse absence rates having a more significant influence on overstaffing variability than larger units (e.g., 30 beds).

Discussion This study highlights the importance of understanding the PNR dynamics in hospital staffing. By deriving the theoretical distribution of PNR and simulating different settings, we approximated the proportion of overstaffed and understaffed days. The results emphasize the sensitivity of PNR to fluctuations in patient volume and nursing availability, underscoring the need for adaptive staffing strategies. This approach allows the evaluation of staffing policies, offering insights for optimizing resource allocation.

10: Real-Time Predictions of Bed Occupancy in Hospitals.

Ensor Rafael Palacios, Theresa Smith

University of Bath, United Kingdom

Increased demand for hospital resources has led to bed occupancy which often approaches and exceeds maximum capacity. Even relatively short periods (e.g., a few days) of elevated bed occupancy can have immediate negative impact at all levels of an hospital service chain, including the number of ambulances available, their response times, and the quality and number of discharges. Predicting periods of high demand, with time horizons up to one or two weeks, is thus of critical operational importance, as it enables hospital managers to proactively initiate adaptive strategies. Here we develop a predictive state-space model of bed occupancy, designed to be deployed within hospitals in real time to support adaptive decision making. We develop and test the model using daily data from two large hospitals in Bristol, United Kingdom. These data include information about bed occupancy itself, admissions, discharges, staffing level and other hospital-level variables; we additionally include information about seasonal infectious diseases (e.g. flu) and weather (e.g., temperature). We benchmark the model against different alternatives, including naive and ARIMA models (with and without covariates) and random forests. For model comparison, we consider multiple loss functions to ensure accurate predictions of different, expert-derived aspects of the data, such as sudden peaks in change occupancy. The next steps involve further validation of the model and testing in an operational setting.

11: Positive and Negative Predictive Values of Diagnostic Tests using Area under the Curve

Kanae Takahashi¹, Kouji Yamamoto²

¹Osaka Metropolitan University Graduate School of Medicine, Japan

²Yokohama City University School of Medicine, Japan

In medicine, diagnostic tests are important for the early detection and treatment of disease. The positive predictive value (PPV) and the negative predictive value (NPV) describe how well a test predicts abnormality. The PPV represents the probability of disease when the diagnostic test result is positive, while the NPV represents the probability of no disease when the diagnostic test result is negative. These predictive values inform clinicians and patients about the probability that the diagnostic test will give the correct diagnosis. Compared to

sensitivity and specificity, the predictive values are more patient focused and often more relevant in patient cases.

However, the predictive values observed in one study do not apply universally because these values depend on the prevalence. In order to overcome the shortcoming, in this study, we proposed a measure of positive and negative predictive values using area under the curve (PPV-AUC and NPV-AUC). In addition, we provided a method for computing confidence intervals of PPV-AUC and NPV-AUC based on the central limit theorem and delta-method.

A simulation study was conducted to investigate the coverage probabilities of the proposed confidence intervals. Simulation results showed that the coverage probabilities of 95% confidence intervals were close to 0.95 when the sample size was large.

12: Freely Accessible Software for Recruitment Prediction and Recruitment Monitoring: Is it Necessary?

Philip Heesen, Manuela Ott, Katarina Zatkova, Malgorzata Roos

University of Zurich, Switzerland

Background

Scientific studies require an adequate number of observations for statistical analyses. The ability of a study to successfully collect the required number of observations ultimately depends on a realistic study design based on accurate recruitment predictions. Inaccurate recruitment predictions inevitably lead to inappropriately designed studies, small sample sizes and unreliable statistical inference, increasing the risk of study discontinuation and wasted funding. To realistically predict recruitment, researchers need free access to statistical methods implemented in user-friendly, well-documented software.

Methods

A recent systematic review assessed the availability of software implementations for predicting and monitoring recruitment.

Results

This systematic review demonstrated that freely accessible software for recruitment predictions is currently difficult to obtain. Although several software implementations exist, only a small fraction is freely accessible. Ultimately, only one article provided a link to directly

applicable free open-source software, but other links were outdated.

Conclusion

To improve access for researchers worldwide, we propose three measures: First, future authors could increase the findability of their software by explicitly mentioning it in titles, abstracts and keywords. Second, they could make their software available online on open access platforms. Finally, they could provide user-friendly documentation and instructive examples on how to use the statistical methods implemented in their software in applications. In the long term, it could become standard practice to use such software for insightful recruitment predictions and realistic decision making. Such realistic decisions would increase the chance that studies are appropriately designed, adequately powered, and successfully completed, thereby optimising the use of limited funding resources and supporting scientific progress worldwide.

13: On Moderation in a Bayesian log-Contrast Compositional Model with a Total. Interaction Between Extreme Temperatures and Pollutants on Mortality

Germá Coenders^{1,2}, Javier Palarea-Albadalejo³, Marc Saez^{1,2}, Maria A. Barceló³

¹Research Group on Statistics, Econometrics and Health (GRECS), University of Girona, Spain

²Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública (CIBERESP). Instituto de Salud Carlos III, Madrid, Spain

³Department of Computer Science, Applied Mathematics and Statistics, University of Girona, Spain

Introduction Compositional regression models with a dependent real variable can be specified as log-contrast models with a zero-sum constraint on the model coefficients. Moreover, the Bayesian approach to model fitting, through the Integrated Nested Laplace Approximation (INLA) method, is gaining increasing popularity to deal with complex data structures such as spatiotemporal observations.

Methods In this work, we combine these elements and extend the approach to encompass both total effects, formally defined in a T-space, and moderation or interaction effects into the data modelling. The interpretation of the results is formulated both, in the original scale of the dependent variable and in terms of elasticities.

An illustrative case study is presented aimed at relating all-cause mortality with the interaction

between extreme temperatures, air pollution composition, and total air pollution in Catalonia, Spain, during the summer of 2022.

Results The results show that extreme temperature, exposure to total pollution and to some pollutants in particular (ozone and particulate matter), allowing for some delay in their effect, were associated with increased risk of dying. Also, again considering delayed effects, the mortality risk particularly increased on days of extreme temperatures and greater exposure to ozone.

Conclusions When assessing the effects of extreme temperatures on mortality, the effects of composition and total pollution, and not just individual pollutants, as well as possible interactions, must be taken into account.

14: Graphical Inference in Nonparametric Hypothesis Testing: A Two-Dimensional Grid Framework for Analysis and Visualization

Lubomír Štěpánek^{1,2}, Ondřej Vít², Lubomír Seif²

¹First Faculty of Medicine of Charles University (Czech Republic)

²Faculty of Informatics and Statistics of Prague University of Economics and Business (Czech Republic)

Background / Introduction Nonparametric tests often utilize intuitive concepts such as ranking of observations or assessing pre-post changes. While these tests, including the Mann-Whitney test and signed-rank tests, offer numerical precision, they can also be interpreted graphically. Though graphical techniques cannot replace numerical calculations, they enhance comprehension of test logic and may lead to practical heuristic formulations.

Methods This study revisits graphical inference testing for selected nonparametric tests, including both two-sample and paired tests. The graphical testing approach transforms test statistic construction into orthogonal directional changes on a two-dimensional finite-step grid. The graphical pathway depends on what the test statistics emphasize in the observations. For two-sample tests, both the ranking distribution and sample affiliation changes matter, whereas, for paired tests, the sequence of positive and negative pre-post changes is critical. These changes are represented as unit steps in orthogonal directions on the grid. Under the null hypothesis of no difference, graphical pathways exhibit an almost regular alternation of grid directions, which follow a binomial distribution and can thus be analyzed within a probability framework. As a novel contribution, we apply Popoviciu's inequality to

derive an upper bound on the probability of observing data contradicting the null hypothesis to the same or a greater extent, thereby estimating the *p*-value and offering insights into the statistical power of the test.

Results We developed R functionality for computing and visualizing two-dimensional grids used in graphical inference testing. The grids highlight regions corresponding to typical null hypothesis rejection scenarios. In particular, the grids accommodate asymmetric null hypotheses for the signed-rank test by upper-bounding directional "traffic" maxima. Various simulations were conducted, evaluating different sample pairs and pre-post scenarios to demonstrate the method's applicability.

Conclusion Graphical inference testing provides an alternative perspective on nonparametric hypothesis testing, fostering better understanding and serving educational purposes. The developed R functionality for graphical testing will soon be integrated into an R package, expanding accessibility and usability for statistical analysis and instruction.

15: On Regression Analysis of Interval-Valued Data Based on Order Statistics

Ryo Mizushima, Asanao Shimokawa

Tokyo University of Science / Japan

Background / Introduction Some of today's diverse data may be given as interval values, such as blood pressure, instead of point values. Interval values can also be used to summarise point-valued data by certain characteristics. For example, the temperature at a certain point in time is given as a point value, but the temperature throughout the day can be described as a minimum and maximum temperature. Most studies in regression analysis of interval-valued data have been proposed based on methods using midpoint and width information. However, those methods rarely consider the information in the interval. In this study, therefore, we consider the case where the upper and lower values of the objective variable and essentially the number of individuals in between are known. An example would be a hospital with 100 patients, where some numerical information on their health status is available, but only the maximum and minimum values are known among them from a privacy point of view.

Methods We propose a model that takes into account the information in the intervals of the objective variable. The proposed method assumes that the values in the interval are generated based on a certain distribution and aims to estimate the distribution of the

objective variable under a given set of explanatory variables. To this end, the upper and lower sides of the objective variable are considered as the maximum and minimum values of the order statistics and the number of observations whose values are not known in the interval is assumed to be known. The maximum and minimum values and the number of observations between them give the conditional probability density function of the objective variable given the explanatory variables from the nature of the order statistic. It is used as a likelihood function to obtain a maximum likelihood estimator of the parameters of the distribution. The estimator can give approximate confidence intervals for the parameters.

Results We checked through simulations the behaviour of parameter estimators and approximate confidence intervals under finite samples when the sample size and the number of objective variables in the interval are varied. The results show that they can be successfully estimated under several conditions.

Conclusion We proposed a method of regression analysis using a likelihood function obtained from the information in the interval of the objective variable, considering the maximum and minimum values of the interval as order statistics.

16: Two-Sided Bayesian Simultaneous Credible Bands in Linear Regression Model

Fei Yang

University of Manchester, United Kingdom

Credible bands, which comprise a series of credible intervals for each component of a parameter vector, are frequently employed to visualize estimation uncertainty in Bayesian statistics. Unlike the often-used pointwise credible interval, simultaneous credible bands (SCBs) can cover the entire parameter vector of interest with an asymptotic probability of at least $1-\alpha$.

In this study, in order to assess where lies the true model $x^T \theta$ from which the observed data have been generated, we propose the two-sided $1-\alpha$ level Bayesian simultaneous credible bands for the regression line $x^T \theta$ over a finite interval of the covariate x in a simple linear regression model. By incorporating the prior information, the proposed method exhibits advantages over the traditional frequentist approach in more robust and stable estimates, especially in cases with limited data.

Using non-informative priors, we analyze the posterior distribution of targeted parameters of interest and employ Monte-Carlo simulations to produce the critical constant related to the construction of Bayesian SCBs.

Simulation results show that the proposed methodology has highly satisfactory frequentist properties. Additionally, it meets the required false-positive rate with a pre-specified level of certainty. Real data analysis in drug stability studies also verify its effectiveness of the proposed framework.

17: Cell Composition Analysis with Unmeasured Confounding

Amber Huybrechts^{1,2}, Koen Van den Berge², Sanne Roels², Oliver Dukes¹

¹Ghent University, Belgium

²Janssen Pharmaceutica, Belgium

Analysis of single-cell sequencing data, in particular cell abundance data where one counts the number of cells detected for each cell type in each sample, involves handling data compositionality. Indeed, cell composition data contain only relative information on a cell type's abundance. An increase in one cell type might therefore also be reflected as a decrease in other cell types' abundance. This makes estimating causal disease effects in cell composition data challenging, especially in the presence of confounders. On top of that, not all confounders might be observed.

Existing methods like *CATE* [1] and *RUV-4* [2] attempt to obtain unbiased disease or treatment effects by estimating the unmeasured confounders using factor analysis and making assumptions on sparsity and the existence of negative controls. However, it is uncertain how these methods perform in the context of cell composition analysis, where in addition to the compositionality, the number of features is smaller in comparison to the settings where these methods are generally used (*e.g. gene expression analysis with thousands of genes*).

In this work, we investigate how we can account for compositionality and unmeasured confounders when assessing differences in cell type abundance between biological conditions. We find that a vanilla factor analysis model, typically used for estimating unmeasured confounders, is unsuitable in the context compositional data, and evaluate alternative approaches.

[1] Jingshu Wang, Qingyuan Zhao, Trevor Hastie, Art B. Owen, "Confounder adjustment in multiple hypothesis testing", *The Annals of Statistics*, *Ann. Statist.* 45(5), 1863-1894, (October 2017)

[2] Johann A. Gagnon-Bartsch, Laurent Jacob, Terence P. Speed "Removing Unwanted Variation from High Dimensional Data with Negative Controls", Berkeley University of California

(December 2013)

18: Temporal Transcriptomic Analysis of Microexon Alternative Splicing in Mouse Neurodevelopmental Genes

Jimin Kim, Kwanghoon Cho, Jahyun Yun, Dayeon Kang

Korea Bioinformation Center (KOBIC), Korea Research Institute of Bioscience and Biotechnology, Korea, Republic of (South Korea)

Background Alternative splicing plays a pivotal role in gene regulation, particularly within neurological processes. Microexons, short exon sequences ranging from 3 to 27 base pairs, are highly neuron-specific and fine-tune protein interactions within synaptic networks. Dysregulation of microexon splicing has been linked to impaired neuronal connectivity and altered synaptic function, which are hallmarks of neurodevelopmental disorders. Understanding the dynamic regulation of microexon splicing across developmental stages is crucial for identifying potential biomarkers and therapeutic targets.

Methods We investigated the temporal dynamics of microexon splicing by analysing whole-cortex RNA sequencing (RNA-seq) data from mice across eleven developmental stages, spanning embryonic, postnatal, and ageing periods. We focused on microexons under 30 base pairs, using the Percent Spliced In (PSI) metric to assess alternative splicing patterns. Our analysis centred on genes involved in neural function and neurodevelopmental disorders to explore the role of microexons in neuronal maturation and synaptic function.

Results We identified distinct stage-specific microexon splicing patterns in several genes, highlighting the complexity of microexon regulation during cortical development. During early embryonic stages (E10–E16), low PSI values were observed for genes involved in neurogenesis and axon guidance, such as Nrcam and Robo1. Nrcam showed a gradual increase in PSI during embryogenesis, whereas Robo1 exhibited a decline from embryonic to postnatal stages, reflecting their roles in neuronal connectivity and circuit stabilisation, respectively. In postnatal stages, Shank3 and Dlgap1 showed significant PSI increases, indicating their involvement in synaptic maturation and plasticity. Conversely, Bin1 displayed a decline in PSI during maturation and ageing, suggesting a shift from synaptic plasticity to stability.

Conclusion This study demonstrates the importance of microexons in neural development and their potential contribution to neurodevelopmental disorders. The stage-specific PSI variations indicate that microexons are crucial for neural circuit formation, synaptic plasticity, and functional specialisation. The observed co-regulation patterns suggest that microexon splic-

ing is tightly regulated, orchestrating key neurodevelopmental events. Future research into the regulatory mechanisms governing microexon splicing will be essential to understanding their broader biological implications and therapeutic potential.

19: A Robust Method for Accurate Reconstruction of 3D Genome Conformation from Hi-C Data

Insu Jang^{1,2}, MInsu Park¹

¹Department of Information and Statistics, Chungnam National University, Korea, Republic of (South Korea)

²Korea Research Institute of Bioscience and Biotechnology, Korea, Republic of (South Korea)

The three-dimensional (3D) organization of the genome within the cell nucleus plays a pivotal role in critical biological processes, including transcriptional regulation, DNA replication, and repair. Disruptions to this spatial organization, such as aberrant chromatin looping or genomic deletions, are linked to various diseases. Despite its significance, resolving the 3D genome architecture has been historically challenging due to the lack of techniques for high-resolution chromatin mapping. The advent of Chromosome Conformation Capture (3C) technologies, particularly Hi-C, revolutionized this field by enabling genome-wide quantification of chromatin interactions. Hi-C produces a contact count map, providing interaction frequencies between genomic loci, which serves as the basis for computational 3D genome reconstruction. However, deriving biologically meaningful 3D structures from Hi-C data remains computationally challenging due to noise and chromatin complexity. To overcome these challenges, we propose a novel, robust methodology combining Thin Plate Spline (TPS) and Non-Metric Multi-Dimensional Scaling (nMDS), specifically designed to infer smooth and biologically plausible 3D genomic structures while being resilient to noise. Our method was rigorously evaluated on simulated datasets encompassing diverse sized structures with varying levels of noise, as well as on real Hi-C data from the IMR90 cell line. Comparative assessments using simulation datasets demonstrated that our approach consistently produced robust and smoother results under varying noise conditions, outperforming existing models in handling varying levels of noise. Furthermore, its predictive validity was substantiated through comparisons with 111 replicate conformations derived from Multiplexed Fluorescence in situ hybridization (M-FISH) imaging, providing strong empirical support for the method and its applications in 3D genome analysis.

20: Balancing Accuracy, Clinical Utility, and Explainability: A Machine Learning Approach to Prostate Cancer Prediction

Luis Mariano Esteban^{1,2}, Rocío Aznar^{1,3}, Angel Borque-Fernando^{4,5}, Alejandro Camón⁵, Patricia Guerrero⁵

¹Escuela Universitaria Politécnica de la Almunia, Universidad de Zaragoza, Spain

²Institute for Biocomputation and Physics of Complex Systems (BIFI), Spain

³Instituto Tecnológico de Aragón, Spain

⁴Department of Urology, Miguel Servet University Hospital, Spain

⁵Health Research Institute of Aragon Foundation, Spain

Background Advances in mathematical modelling have significantly improved cancer diagnosis. While these models enhance predictive performance, typically measured by discriminative power, they often overlook their role as classification tools. Recently, greater emphasis has been placed on their clinical utility and explainability, highlighting the need for models that balance accuracy with interpretability. Tools such as clinical utility curves and Shapley values can help achieve this balance.

Methodology and results

We analysed data from 86,359 patients at Miguel Servet University Hospital, Zaragoza, Spain (2017–2022) with at least one PSA measurement, including 2,391 prostate cancer diagnoses, to develop a predictive model for PCa. From their clinical records, we selected approximately 50 demographic and clinical variables as candidate predictors, including PSA, free PSA, PSA history, blood analysis parameters, and comorbidities. Several machine learning models were tested, including logistic regression, ridge regression, LASSO, elastic net, classification trees, random forest, neural networks, and Extreme Gradient Boosting (XGBoost). Model performance was validated using an external dataset of 47,284 patients from the Lozano Blesa University Hospital.

XGBoost demonstrated the best discrimination in the validation cohort, with an AUC of 0.965, sensitivity of 0.904, and specificity of 0.914. More importantly, it also showed the highest clinical utility. For a cutoff that resulted in a 5% diagnostic loss in the training dataset, the validation dataset showed a 7.87% loss while recommending biopsy for 11.1% of patients. In comparison, a screening policy of biopsying all patients with $\text{PSA} > 3$ would result in 15.3%.

To assess variable influence within the XGBoost model, we used SHAP values (SHapley Additive exPlanations), a game theory-based method for evaluating feature importance in predictive models. SHAP values indicate the contribution of each variable for each individual

and can be analysed collectively or individually. In our analysis, PSA was the most influential risk factor, producing the highest Shapley values. Protective factors included older age, multiple PSA readings between 3.2 and 8 with negative biopsies, and prolonged use of antihypertensives, statins, or antidiabetics. Conversely, a previous negative biopsy with ASAP or PIN was a notable risk factor.

Conclusions This study developed a predictive tool for prostate cancer with high accuracy while minimising unnecessary biopsies. As screening protocols remain unstandardised in Spain, it is crucial to explore alternative strategies that incorporate models capable of reflecting variable importance and clinical utility before implementation.

21: Development of a Thyroid Cancer Recurrence Prediction Calculator: A Regression Approach

Jiaxu Zeng

University of Otago, New Zealand

Background The thyroid cancer staging calculator has been recognised as one of the most efficient tools for assisting clinicians in making clinical treatment decisions. However, the current calculator is missing patients' serum Thyroglobulin information, which is crucial for staging cancer patients in practice. The primary aim of this study is to update current calculator with serum thyroglobulin included based on the tertiary thyroid cancer service database from Australia.

Methods Records from 3962 thyroid patients were analysed for training a logistic model for predicting recurrence. Twelve predictive variables were chosen under close guidance of thyroid cancer specialists, which includes age at operation, sex, number of carcinomas presented in the operation, size of the greatest tumour, histologic type of carcinoma, extrathyroidal extension status of tumours, pathologic staging of the primary tumour, presence of venous invasion of the primary tumour, immunohistochemistry for the primary tumour, presence of extranodal spread, number of lymph nodes and serum thyroglobulin level presented in the scans.

Results The strongest predictors were number of lymph nodes, histologic type of carcinoma and most importantly, the serum thyroglobulin level. The model demonstrated excellent performance with an AUC of 0.874.

Conclusions This study has addressed an important concern that serum thyroglobulin in-

formation was not used to predict thyroid cancer recurrent in practice.

22: A Comparison of Methods for Modelling Multi-State Cancer Progression using Screening Data with Censoring after Intervention

Eddymurphy U. Akwiwu, Veerle M.H. Coupé, Johannes Berkhof, Thomas Klausch

Amsterdam UMC, Amsterdam, The Netherlands

Background Optimizing cancer screening and surveillance frequency requires accurate information on parameters such as sojourn time and cancer risk from pre-malignant lesions. These parameters can be estimated using multi-state cancer models applied to screening or surveillance data. Although multi-state model methods exist, their performance has not been thoroughly investigated, specifically not in the common setting where cancer precursors are treated upon detection so that the transition to cancer is prevented. Our main goal is understanding the performance of available multi-state methods in this challenging censoring setting.

Methods Six methods implemented in R software packages (*msm*, *msm* with a phase-type model, *cthmm*, *smms*, *BayesTSM*, and *hmm*) were compared. We assumed commonly used time-independent (i.e., exponential) or time-dependent (i.e., Weibull) progression hazards between consecutive health states in a three-state model (healthy, HE; cancer precursor; cancer) in simulation studies. Bias, empirical standard error (ESE), and root mean squared error (rMSE) of progression risk estimates were compared across methods. The methods were illustrated using surveillance data from 734 individuals at increased risk of colorectal cancer, classified into three health states: HE, non-advanced adenoma (nAA), and advanced neoplasia (AN). Age was used as the time scale in the analysis, with both the risks estimates of developing nAA from HE and AN after the onset of nAA compared across the methods.

Results All methods performed well with time-independent progression hazards in simulation study. With time-dependent hazards, only the packages *smms* and *BayesTSM* provided unbiased risk estimates with low ESE and rMSE. In the application (median follow-up: 6 years), 447 (65%), 208 (28.3%) and 49 (6.7%) individuals were classified as HE, nAA and AN, respectively. Only the packages *msm*, *hmm*, and *BayesTSM* yielded converged solutions. The risks estimates of developing nAA from HE were similar between *hmm* and *BayesTSM* (e.g., nAA risks estimates at age 30 were approximately zero to 2 decimal places) but differed for the *msm* package (e.g., nAA risk estimate at age 30 was 16%), while the risks estimates of developing AN after the onset of nAA varied (5-year risk range: 3% to 23%) across methods.

Conclusion Methods for multi-state cancer models, more specifically with unobservable precursor to cancer transition, are strongly impacted by the time dependency of hazard. Careful consideration is crucial when selecting a method for multi-state cancer models. With more realistic (time-dependent hazard) models, the BayesTSM and smms packages performed accurately. However, BayesTSM outperformed in situations with weakly identifiable likelihoods.

23: ADHD and 10-year Disease Progression from Initiating Pharmacotherapy for Hypertension to Death: A Multistate Modelling Analysis

Yiling Zhou¹, Douwe Postmus¹, Anne van Lammeren², Casper F.M. Franssen¹, Harold Snieder¹, Catharina A. Hartman¹

¹University of Groningen, University Medical Center Groningen, Netherlands, The

²Expertisecentrum Fier, Leeuwarden, the Netherlands.

Background Attention-deficit/hyperactivity disorder (ADHD)—the most common neurodevelopmental disorder—affects 2.5% of adults globally and is associated with a 1.5–2-fold increased risk of hypertension, which typically manifests a decade earlier than in the general population. However, this population regarding their cardiovascular health has been largely overlooked in research and clinical practice. This nationwide cohort study aims to investigate 10-year disease trajectories after initiating hypertension pharmacotherapy in adults with ADHD using multistate modelling.

Methods This nationwide cohort study included adults aged 18–90 years in the Netherlands who initiated hypertension medication between 2013 and 2020, without prior cardiovascular disease (CVD) or chronic kidney disease (CKD). Hypertension was defined as the initial state, critical complications (stroke, heart failure hospitalisation [HHF], acute myocardial infarction, and CKD) as intermediate states, and death from a cardiovascular or renal cause (cardiorenal death) or other causes as final states. Transition rates were estimated using Cox proportional hazards regression, individual-level trajectories were generated via microsimulation, and the effect of ADHD was estimated by comparing outcomes in individuals with ADHD to a counterfactual scenario where individuals were assumed not to have ADHD.

Results Of 592,362 adults included, 9,728 had ADHD (median age, 45.0 years). Compared to the counterfactual scenario, individuals with ADHD had a higher 10-year risk of cardiorenal death via the HHF pathway (risk difference [95% CI]: 4.8 [2.0–9.2] per 10,000 persons), driven by increased transition risks from hypertension to HHF (14.2 [7.6–26.1] per 10,000 persons), and HHF to cardiorenal death (752.8 [55.7–1517.9] per 10,000 persons). Similarly,

individuals with ADHD had an elevated 10-year risk of cardiorenal death via the CKD pathway (2.8 [1.3–7.0] per 10,000 persons), primarily due to an increased transition risk from CKD to cardiorenal death after CKD onset (351.6 [175.7–782.4] per 10,000 persons).

Conclusion In individuals initiating hypertension medication without pre-existing CVD or CKD, ADHD was associated with a worse 10-year prognosis of hypertension, particularly for the pathways initiated by heart failure and CKD. Our findings indicate the importance of interdisciplinary care and highlight the need for research aimed at preventing heart failure after hypertension onset and optimising heart failure and CKD management in individuals with ADHD.

24: Understanding PSA Dynamics: Integrating Longitudinal Trajectories, Testing Patterns, and Disease Progression.

Birzhan Akynkozhayev, Benjamin Christoffersen, Mark Clements

Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Sweden

Background

The prostate-specific antigen (PSA) test is a widely used, inexpensive test for prostate cancer screening and prognosis. However, its clinical utility remains debated. Examining PSA trajectories over time provides deeper insights into prostate cancer risk and progression. We believe that a critical challenge in such analyses is the observational process: men with higher PSA levels tend to undergo more frequent testing, introducing bias when PSA trajectories are evaluated without adjustment for different follow-up patterns. This study utilises the Stockholm Prostate Cancer Diagnostics Register, which contains PSA measurements from over half a million men living in Stockholm between 2003 and 2023.

Methods

Longitudinal mixed-effects models were applied to characterise the PSA trajectories. Separate survival models were fitted for time-to-prostate-cancer-diagnosis and the observational process, where time-to-next PSA test was treated as a recurrent event. A full joint model was fitted to incorporate both processes, examining different association structures---including current PSA value, rate of change, and cumulative PSA levels---for their predictive impact on prostate cancer diagnosis and testing behaviour. The survival component of the model incorporated recurrent events (time-to-next-PSA test for the observational process) alongside a terminal event (time-to-diagnosis) while also accounting for delayed entry. Model estimation was facilitated by our recently developed VAJointSurv framework, allowing scalable inference through variational approximations for fast integrations.

Results

Our findings highlight the importance of accounting for the observational process in PSA testing. Frequent follow-up testing among men with higher PSA values influenced PSA trajectory characterisation and hazard estimates for diagnosis. We found that modelling the observational process and disease progression separately yielded different results compared to a joint approach, which combines and accounts for both processes. These findings indicate that joint modelling may provide a more comprehensive understanding of PSA dynamics and its relationship with disease progression, rather than modelling each process separately.

Conclusions By jointly modelling PSA trajectories, disease progression, and the observational process, we provide a robust framework for understanding the relationship between these related processes. To our knowledge, this is the largest study of longitudinal and joint PSA modelling. These findings may aid researchers who are exploring PSA trajectories over time. This approach highlights the necessity of adjusting for the observational process to derive accurate assessments of PSA trajectories and prostate cancer risk.

25: Joint Modeling for Principal Stratification: Analyzing Stroke Events in CADASIL Patients Across NOTCH3 Variants

Léa Aguilhon, Sophie Tezenas du Montcel, Juliette Ortholand

Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, F-75013, Paris, France, France

Background Principal Stratification is a statistical framework designed to analyze causal effects by considering subgroups of individuals defined by their potential outcomes, often in the context of mediators or intermediate variables. This method is especially valuable for addressing challenges where treatment effects are influenced by intermediate events, such as competing events, allowing for a more accurate estimation of causal effects. While powerful, principal stratification relies on unobservable counterfactual strata, which raises identifiability challenges, often requiring strong assumptions too. Recent methodological advancements have focused on reducing these assumptions with estimation techniques. In parallel, powerful estimators such as joint models have been developed as effective tools for predicting event outcomes using repeated measures.

Objective This study uses joint modeling to reduce reliance on untestable assumptions in principal stratification. We applied this approach to study Cerebral Autosomal Dominant Arteriopathy with Subcortical Infarcts and Leukoencephalopathy (CADASIL), a genetic disorder that impacts small blood vessels. We analyze stroke occurrence in the presence of

death across NOTCH3 variants (1-6 and 7-34).

Method We analyzed observational follow-up data from 337 CADASIL patients who were followed on average for 5.6 years. We studied the occurrence of the second stroke event (83 observed events) in the context of death truncation (70 observed events) and used functional and cognitive scores to inform both events' occurrences.

Membership of the main strata was obtained from the predicted counterfactual of death from a Bayesian multivariate joint model (JMBayes2 package). Inverse Probability of Treatment Weighting was used to adjust on covariates as sex, cardiovascular risks, education level, and baseline scores. And Restricted Mean Survival Time (RMST) was then applied to quantify stroke-free survival in the "always-survivor" subpopulation.

Results Preliminary analysis indicates that carriers of the NOTCH3 1-6 variant have a lower 2-year and 5-year restricted mean stroke-free survival time compared to non-carriers, suggesting an accelerated time to second stroke.

Conclusion We used joint modeling to estimate the probability of each patient of belonging to the principal strata (always survivor), in context of death truncation. This permits alleviating assumptions necessary for Principal Stratification estimation. Finally, this study provides valuable insights into CADASIL progression and paves the way for the design of future clinical trials.

26: Challenges in Modelling Neuropsychiatric Symptoms in Early Alzheimer's Disease

Rachid Abbas

F. Hoffman -La Roche, France

The Neuropsychiatric Inventory (NPI) is a structured caregiver-based questionnaire designed to assess and quantify neuropsychiatric symptoms in individuals with various neurodegenerative disorders. NPI is a widely utilized instrument in clinical and research settings to provide a comprehensive evaluation of behavioral and psychological symptoms associated with cognitive impairment. The NPI has demonstrated good reliability and validity across various neurological conditions, such as Alzheimer's disease (AD), frontotemporal dementia, and vascular dementia.

Most of the current AD clinical trials are focused on the early stage of the disease when

neuropsychiatric symptoms are rare, but there is a growing interest in the detection of their incidence as this may be viewed as a clinically meaningful hallmark of deterioration of the quality of life. From an analytical perspective, this requires to analyse a continuous variable with an excess of 0, also described as overdispersed data.

Over-dispersed data refers to a situation where the variance of observed data exceeds what would be expected under a theoretical distribution, such as a Poisson distribution. This departure from the assumptions of homogeneity in variance poses challenges in statistical modeling, as it may lead to inefficient parameter estimates and inflated Type I error rates. Many analytical solutions were proposed to tackle such issues and contribute to a more accurate and robust statistical analysis of over-dispersed count data.

In this work, we quantified the benefits in terms of predictive performance and type I error control of various analytical approaches to handle over-dispersed NPI data. Our findings allow us to make evidence-based recommendations on analysis strategies. By optimizing the statistical approach to NPI data analysis, we pave the way for more sensitive and reliable detection of treatment effects on neuropsychiatric symptoms in early-stage AD clinical trials. As we continue to push the boundaries of AD drug development, these methodological advancements will be crucial in unlocking new possibilities for upcoming targeted interventions.

27: Network Models to Decipher the Human Exposome : Application to Food Exposome Patterns in the General Population

Ima Bernada¹, Gregory Nuel², Cécilia Samieri¹

¹INSERM U1219, France

²LPSM, CNRS 8001

Complex chronic diseases are partly due to the exposome. Some exposures co-occur in usual life and combination of exposures, rather than single factors, contribute to disease development. In co-exposure modeling, most studies use risk scores or dimension reduction approaches. These ignore important features of the dependency structure, like highly connected variables that may play a central role in disease development. Network approaches allow to capture the full complexity of the exposome structure. Our objective was to decipher the food exposome, encompassing both intakes and biological fingerprints, in a large cohort of older persons. We aimed at characterizing diet intake networks, understanding how they may be reflected internally through diet-related metabolites networks, and integrating the two in a bipartite network.

We analyzed a sample of n=311 participants from the 3C-Bordeaux cohort study who answered a dietary survey with assessment of intakes in 32 food groups (n=1730) and provided blood draw for measurement of 143 food-related metabolites (n=375). Using MIIC algorithm based on conditional mutual information, we constructed three co-exposure networks: (i) food co-consumption network; (ii) food-related metabolite network; and (iii) food-to-metabolite bipartite network. To address estimation uncertainty, networks were analyzed through bootstrap replication, using graph theory metrics (e.g. degrees, distances). Obtaining collections of networks by bootstrap replication enabled to quantify the uncertainty of each link. This approach allowed a rigorous analysis of the results. A consensus network was also searched and represented. The networks were also studied through clustering, on one hand by using a priori clusters (e.g. metabolite families), and on the other hand using a node clustering method.

The consensus food co-consumption network reflected the French southwest diets of older person living in Bordeaux in early 2000. A subnetwork centered on potatoes indicated that its consumption was central and closely linked to that of many other foods contained in the traditional south-western diet. The metabolite network showed expected links between metabolites originating from the same food sources or the same biological pathways. Finally, when taking an interest in the links between food components and metabolites, we found expected biological links and more novel associations which warrant further investigations.

Network approaches applied simultaneously to food intakes and food-derived metabolites allow to integrate both external and internal parts of the food exposome in a single statistical framework. This integrated behavioral-biological approach gives novel insights on how environmental exposures such as diet impact biology and health.

28: Study Brain Connectivity Changes in Dementia with Lewy Bodies with Functional Conditional Gaussian Graphical Models

Alessia Mapelli^{1,2}, Laura Carini³, Michela Carlotta Massi², Dario Arnaldi^{4,5}, Emanuele Di Angelantonio^{2,6,7}, Francesca Ieva^{1,2}, Sara Sommariva³

¹MOX – Laboratory for Modeling and Scientific Computing, Department of Mathematics, Politecnico di Milano, Italy

²HDS– Health Data Science Center, Human Technopole, Italy

³Università degli studi di Genova, Department of Mathematics, Italy

⁴Department of Neuroscience, Rehabilitation, Ophthalmology, Genetics, Maternal and Child Health (DINOGMI), University of Genoa, Genoa, Italy

⁵Clinical Neurophysiology Unit, IRCCS Ospedale Policlinico S. Martino, Genoa, Italy

⁶Blood and Transplant Research Unit in Donor Health and Behaviour, Cambridge, UK

⁷Dept of Public Health & Primary Care, University of Cambridge, Cambridge, UK

Dementia with Lewy Bodies (DLB) represents the second most common cause of neurodegenerative dementia after Alzheimer's Disease. Alterations in functional connectivity in the brain are possible phenotypic expressions of this disorder. Multivariate time series analysis of signals resulting from electroencephalography (EEG) is extensively used to study the associations between simultaneously recorded signals, and thus quantify functional connectivity at a subject level. Network-Based Statistic is a modern approach used to perform statistical group-level analysis and identify differential connectivity graphs between groups of patients presenting different clinical features . However, current methods fail to distinguish between direct and indirect associations between brain areas and often neglect the impact of confounding factors. We propose a conditional Gaussian graphical model for multivariate random functions to achieve a population-level representation of the conditional dependence of brain functionality captured by EEG, allowing the graph structure to vary with the external variables.

Our method builds on the work of Zhao et al. [1], extending their high-dimensional functional graphical model to account for external variables. In this approach, each node in the graph represents the signal from an EEG electrode. We adopt a neighborhood selection strategy to estimate sparse brain connectivity graphs based on penalized function-on-function regression. Briefly, each node's signal is predicted from the signals of all other nodes using a lasso-based function-on-function regression. External variables (such as phenotype and age) are included in the model as interactions with the signals to capture their influence on brain connectivity. By combining the estimated neighborhoods, we recover the complete graph structure, which can adapt to variations in external factors. The key advantage of this method is its ability to detect differential connectivity changes associated with specific conditions modeling confounder-linked networks for more accurate estimates.

The method was first validated through simulated data mimicking high-density EEG data recorded using a 64-electrode cap during an eyes-closed resting state task. The method was then tested on experimental data demonstrating the capability of the proposed approach in characterizing differences in functional connectivity in DLB patients with different clinical features, including hallucinations, fluctuations, parkinsonism, and REM sleep behavior disorder.

This study introduces a novel conditional graphical model for multivariate random functions that enables more precise modeling of brain connectivity by accounting for conditional relationships and mitigating confounding bias in differential network analysis.

[1] Zhao, Boxin, et al. "High-dimensional functional graphical model structure learning via neighborhood selection approach." *Electronic Journal of Statistics* 18.1 (2024): 1042-1129.

29: LongiSurvSHAP: Explaining Survival Models with Longitudinal Features

Van Tuan Nguyen, Lucas Ducrot, Agathe Guilloux

Inria, Université Paris Cité, Inserm, HeKA, F-75015 Paris, France

Background Recent developments in survival models integrating longitudinal measurements have significantly improved prognostic algorithm performance (Lee, Yoon, and Van Der Schaar 2019; Bleistein et al. 2024). However, their complexity often renders them black boxes, limiting applicability, particularly in critical fields like healthcare. Regulatory frameworks in the EU and the US now require interpretability tools to ensure model predictions align with expert reasoning, thereby enhancing reliability (Geller 2023; Panigutti et al. 2023). Despite this requirement, research on explaining these models remains limited, and existing methods are often constrained to specific architectures (Lee, Yoon, and Van Der Schaar 2019).

Methods We introduce LongiSurvSHAP, a model-agnostic explanation algorithm designed to interpret any prognostic model based on longitudinal data. While TimeSHAP (Bento et al. 2021) extends the concept of SHapley Additive exPlanations (SHAP) to time series classification, we advance this framework to survival analysis, accommodating irregular measurements of longitudinal features, which is ubiquitous in healthcare.

Results Our algorithm provides both individual and global explanations. Extensive simulations demonstrate LongiSurvSHAP's effectiveness in detecting key features and identifying crucial time intervals influencing prognosis. Applied to data from MIMIC (Johnson et al. 2016), our method aligns with established clinical knowledge, confirming its utility in real-world healthcare scenarios.

Conclusion We present a novel algorithm that enhances interpretability in survival analysis by revealing the impact of longitudinal features on survival outcomes.

References Bento, João, Pedro Saleiro, André F Cruz, Mário AT Figueiredo, and Pedro Bizarro (2021). “Timeshap: Explaining recurrent models through sequence perturbations”. In: Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining, pp. 2565–2573.

Bleistein, Linus, Van-Tuan Nguyen, Adeline Fermanian, and Agathe Guilloux (2024). “Dynamic Survival Analysis with Controlled Latent States”. In: Forty-first International Conference on Machine Learning.

Geller, Jay (2023). "Food and Drug Administration Published Final Guidance on Clinical Decision Support Software". In: *Journal of Clinical Engineering* 48.1, pp. 3–7.

Johnson, Alistair EW et al. (2016). "MIMIC-III, a freely accessible critical care database". In: *Scientific data* 3.1, pp. 1–9.

Lee, Changhee, Jinsung Yoon, and Mihaela Van Der Schaar (2019). "Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data". In: *IEEE Transactions on Biomedical Engineering* 67.1, pp. 122–133.

Panigutti, Cecilia et al. (2023). "The role of explainable AI in the context of the AI Act". In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1139–1150.

30: Prognostic Models for Recurrent Event Data

Victoria Watson^{1,2}, Laura Bonnett², Catrin Tudur-Smith²

¹Phastar, United Kingdom

²University of Liverpool, Department of Health Data Sciences

Background / Introduction Prognostic models predict outcome for people with an underlying medical condition. Many conditions are typified by recurrent events such as seizures in epilepsy. Prognostic models for recurrent events can be utilised to predict individual patient risk of disease recurrence or outcome at certain time points.

Methods for analysing recurrent event data are not widely known or applied in research. Most analyses use survival analysis to consider time until the first event, meaning subsequent events are not analysed and key information is lost. An alternative is to analyse the event count using Poisson or Negative Binomial regression. However, this ignores the timing of events. Recurrent event methods analyse both the event count and the timing between events meaning key information is not discarded.

Methods A systematic review on methodology for analysing recurrent event data in prognostic models was conducted. Results from this review identified methods commonly used in practice to analyse recurrent event data. A simulation study was then conducted which evaluated the most frequently identified methods in the systematic review with respect to the underlying event rate. The event rates were categorised into low, medium and high based on data collected in the systematic review to best represent a variety of chronic conditions

or illnesses where recurrent events are typically seen.

Results The simulation study provided evidence to determine if model choice may be influenced by the underlying event rate in the data. This was assessed by deriving statistics suitable for recurrent event methods to assess the model fit and predictive performance of the recurrent event methods. These statistics were used to determine if certain methods identified tended to perform better than others under different scenarios.

Conclusion Results from the systematic review and simulation study will be presented including a summary of each method identified. The results will be the first step towards a toolkit for future analysis of recurrent event data.

31: Unlocking Diagnosis Code for Longitudinal Modeling Through Representations from Large Language Models

Fabian Kabus¹, Maren Hackenberg¹, Moritz Hess¹, Simon Ging³, Maryam Farhadizadeh², Nadine Binder², Harald Binder¹

¹Institute of Medical Biometry and Statistics (IMBI), Medical Center, University of Freiburg

²Institute of General Practice/Family Medicine, Medical Center, University of Freiburg

³Department of Computer Science, Faculty of Engineering, University of Freiburg

Background In longitudinal data, there often is a multitude of diagnosis codes, such as ICD-10 codes, in particular when considering clinical routine data. Incorporating a large number of codes can be challenging, as treating them as categorical variables in statistical models leads to a large number of parameters, and also one-hot encoding, often used in machine learning, provides no solution to this. In addition, the actual meaning of the diagnoses is not captured. There, large language models might provide a solution, as they capture meaning and can provide alternative numerical representations via their embeddings. We consider such an approach specifically in the context of longitudinal modeling with transformer neural networks.

Methods We generate embeddings using pre-trained language models and refine them during training in the longitudinal prediction task. Specifically, we compare two embedding strategies, sentence embeddings from SBERT and attention-weighted pooled hidden states from LLaMa, with one-hot encoding as a baseline. Additionally, we investigate different text generation strategies, using either standard ICD-10 descriptions or expanded descriptions generated via prompt-engineered large language models. To evaluate the structure of the learned embeddings, we apply TriMap for dimensionality reduction, assessing whether

language-based embeddings capture more coherent relationships between ICD-10 codes.

Results On a clinical routine dataset, models initialized with language-based embeddings derived from sentence-level representations outperform the one-hot encoding baseline in prediction performance, while embeddings extracted from the larger autoregressive model do not show a consistent improvement. Visualization using TriMap suggests that sentence-level embeddings lead to more coherent clustering of ICD-10 codes, capturing their semantic relationships more effectively. Attention analysis indicates that the transformer utilizes these structured embeddings to enhance prediction performance. Additionally, results suggest that incorporating domain-specific prompt engineering further refines embedding quality, leading to more distinct and clinically informative code representations.

Conclusion Integrating textual descriptions into ICD-10 embeddings enhances prediction modeling by providing a structured initialization that incorporates domain knowledge upfront. As large language models continue to evolve, this approach allows advancements in language understanding to be leveraged for longitudinal medical modeling.

32: Machine Learning Perspectives in Survival Prediction Model Selection: Frequentist vs. Bayesian Approach

Emanuele Koumantakis¹, Valentina Bonuomo¹, Selene Grano², Fausto Castagnetti³, Carlo Gambacorti-Passerini⁴, Massimo Breccia⁵, Maria Cristina Miggiano⁶, Chiara Elena⁷, Matteo Pacilli⁸, Isabella Capodanno⁹, Tamara Intermesoli¹⁰, Monica Bocchia¹¹, Alessandra Iurlo¹², Fabio Ciceri¹³, Fabrizio Pane¹⁴, Federica Sorà¹⁵, Barbara Scappini¹⁶, Angelo Michele Carella¹⁷, Elisabetta Abruzzese¹⁸, Sara Galimberti¹⁹, Sabrina Leonetti Crescenzi²⁰, Marco de Gobbi¹, Giuseppe Saglio¹, Daniela Cilloni¹, Carmen Fava¹, Paola Berchialla¹

¹Department of Clinical and Biological Sciences, University of Torino, Torino, Italy

²Department of Molecular Biotechnologies and Health Sciences, University of Torino, Torino, Italy

³Department of Medical and Surgical Sciences, Institute of Hematology "Seragnoli", University of Bologna, IRCCS Azienda Ospedaliero-Universitaria di Bologna, Bologna, Italy

⁴Department of Medicine and Surgery, University Milano-Bicocca, Monza, Italy

⁵Department of Translational and Precision Medicine, Az. Policlinico Umberto I-Sapienza University, Rome, Italy

⁶Hematology Department, San Bortolo Hospital, Vicenza U.O.C. di Ematologia, Vicenza, Italy

⁷U.O.C. Ematologia 1, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy

Abstracts of Contributed Posters

⁸U.O.C. Ematologia, Grande Ospedale Metropolitano Bianchi-Melacrino-Morelli, Reggio Calabria, Italy

⁹Hematology, AUSL Reggio Emilia, Reggio Emilia, Italy

¹⁰Hematology and Bone Marrow Transplant Unit, Azienda Socio-Sanitaria Regionale Papa Giovanni XXIII, Bergamo, Italy

¹¹Hematology Unit, Azienda Ospedaliera Universitaria Senese, University of Siena, Siena, Italy

¹²Hematology Division, Foundation IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy

¹³Hematology and Bone Marrow Transplantation Unit, IRCCS San Raffaele Hospital, Milan, Italy

¹⁴Hematology and Hematopoietic Stem Cell Transplant Center, Department of Medicine and Surgery, University of Naples Federico II, Naples, Italy

¹⁵Fondazione Policlinico Universitario Agostino Gemelli IRCCS, Rome, Italy

¹⁶Hematology Unit, Azienda Ospedaliero-Universitaria Careggi, Florence, Italy

¹⁷Hematology and Bone Marrow Transplant Unit, IRCCS Fondazione Casa Sollievo della Sofferenza San Giovanni Rotondo, Foggia, Italy

¹⁸Department of Hematology S. Eugenio Hospital, Rome, Italy

¹⁹Department of Clinical and Experimental Medicine, University of Pisa, Pisa, Italy

²⁰Division of Hematology, Azienda Ospedaliera San Giovanni Addolorato, Rome, Italy

INTRODUCTION

Predictive model selection remains one of the most challenging and critical tasks in medical statistics, particularly in survival analysis or high-dimensional prediction settings. The Cox proportional hazards model is widely used for its simplicity and interpretability but struggles with high-dimensional data, multicollinearity, and overfitting [1]. Stepwise selection methods, while intuitive, suffer from instability, inflated type I error rates, and a tendency to produce overly optimistic models due to their reliance on multiple hypothesis testing. Alternatives like adaptive Lasso and Bayesian Model Averaging (BMA) incorporate regularization and probabilistic frameworks to improve model performance [2,3]. This study focuses on identifying predictive factors for treatment restart in patients who discontinued tyrosine kinase inhibitor (TKI) therapy, using data from the Italy-TFR longitudinal study.

METHODS

The Italy-TFR study is a multicenter observational study evaluating treatment-free remission (TFR) feasibility in chronic myeloid leukemia (CML). We included patients who achieved deep molecular response, discontinued TKI, and had at least one year of follow-up. Survival analysis considered time from TKI discontinuation to restart or last follow-up. Since different model selection strategies can yield different results, we compared Cox proportional hazards model including the whole set of predictors (considered as baseline model), bidirectional step-

wise model selection, Multimodel Inference (MMI), adaptive Lasso, and BMA as predictive models of the risk of restarting treatment.

RESULTS

Among 542 patients from 38 centers, the predictive value of nine independent variables was analyzed. MMI, adaptive Lasso, and BMA identified TKI treatment duration as the most significant predictor of treatment resumption. Stepwise regression, in contrast, selected three variables: duration of therapy, generation of last TKI discontinued, and Sokal Score. The Bayesian Information Criterion (BIC) was lower for MMI, adaptive Lasso, and BMA (2131.242) compared to stepwise regression (2137.84), suggesting better model performance.

CONCLUSIONS

MMI, Adaptive Lasso, and BMA outperformed stepwise regression based on BIC, identifying TKI treatment duration as the most significant predictor. These findings show the advantages of regularization and probabilistic frameworks in improving model stability and interpretability, highlighting their great potential for predictive modeling.

REFERENCES

1. Babyak, M. A. (2004). What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine*, 66(3), 411-421.
2. Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476), 1418-1429.
3. Hoeting et al. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, 14(4), 382-417.

33: Non-Parametric Methods for Comparing Survival Functions with Censored Data: Exhaustive Simulation of all Possible Beyond-Observed Censoring Scenarios and Computational Analysis

Lubomír Štěpánek^{1,2}, Ondřej Vít², Lubomír Seif²

¹First Faculty of Medicine of Charles University (Czech Republic)

²Faculty of Informatics and Statistics of Prague University of Economics and Business (Czech Republic)

Background / Introduction Comparing survival functions, which describe the probability of not experiencing an event by a given time in two groups, is one of the fundamental tasks in survival analysis. Standard methods, such as the log-rank test, Wilcoxon test, and score-rank test of Cox's proportional hazards model and its variants, may rely on statistical assumptions, including sufficient sample size for asymptotic validity or even proportional hazards. However, these assumptions may not always hold, limiting their applicability. This study introduces a non-parametric alternative for comparing survival functions that minimizes assumptions and offers a direct computation of the *p*-value.

Methods Unlike traditional approaches requiring hazard function estimation, our method models all possible scenarios based on observed data, encompassing cases where survival functions differ at least as much as observed. This exhaustive scenario-based modeling enables direct *p*-value calculation without reliance on asymptotic approximations. Given that censoring introduces additional uncertainty, we address its impact by considering a comprehensive (and often large) set of all potential survival function differences. Due to the computational intensity of enumerating all scenarios (coming from observed censoring), we compare a fully exhaustive computational approach with a Monte Carlo simulation-based method. The performance of these approaches is evaluated against the log-rank test, particularly in terms of Type I error rate and computational efficiency. Additionally, we analyze the asymptotic time complexity of both proposed approaches.

Results Based on simulation outputs, our method reduces the Type I error rate compared to the log-rank test, making it particularly useful in settings requiring robustness against false positives. The exhaustive approach ensures an exact *p*-value calculation but is computationally demanding. The Monte Carlo-based approximation significantly improves computational efficiency while maintaining acceptable accuracy, making it a viable alternative for large datasets. Our complexity analysis highlights the trade-offs between computational cost and statistical precision.

Conclusion The proposed non-parametric method provides an alternative to traditional survival function comparison techniques. A novel aspect of our approach is the calculation of all possible scenarios for censored observations when estimating the counts of survival functions that are at least as different as observed. By directly evaluating all plausible scenarios, it reduces reliance on assumptions while improving Type I error rate control. The Monte Carlo approximation offers a computationally feasible alternative, retaining statistical robustness in

practical applications. These findings support the use of assumption-minimized approaches in survival analysis, particularly in studies where conventional methods may be restrictive.

34: Tree-Based Methods for Length-Biased Survival Data

Jinwoo Lee¹, Jiyu Sun¹, Donghwan Lee²

¹Integrated Biostatistics Branch, National Cancer Center, Republic of Korea

²Department of Statistics, Ewha Womans University, Republic of Korea

Background Left truncation in prevalent cohort studies, where only individuals who have experienced an initiating event (such as disease onset) and survived until study enrollment are observed, leads to length-biased data when the onset follows a stationary Poisson process. Although the existing survival trees and survival forests for left-truncated right-censored (LTRC) data can be applied to estimate survival functions, they may be inefficient for analyzing length-biased right-censored (LBRC) data.

Methods We proposed tree-based methods for LBRC data by adapting the conditional inference tree (CIT) and forest (CIF) frameworks. Unlike LTRC-based approaches, which use log-rank scores from a conditional likelihood, our methods employed log-rank scores derived from the full likelihood, which is valid under LBRC settings. To improve numerical stability and computational efficiency, we adopted a closed-form cumulative hazard function (CHF) estimator for log-rank scores as an alternative to the nonparametric maximum likelihood estimator.

Results Simulation studies indicated that LBRC-CIT achieves a higher recovery rate of the true tree structure in LBRC data than conventional LTRC-CIT, with particularly notable benefits in small-sample settings. Under proportional hazards and complex nonlinear LBRC scenarios, LBRC-CIF offers more accurate predictions than LTRC-CIF. We illustrated the application of our methods to the estimation of survivorship using a dataset of lung cancer patients with COPD.

Conclusions By using full-likelihood-based log-rank scores and a closed-form CHF estimator, our proposed LBRC-CIT and LBRC-CIF methods enhance both statistical efficiency and computational stability for length-biased right-censored data.

35: Evaluating Different Pragmatic Approaches for Selecting the Truncation Time of the Restricted Mean Survival Time in Randomized Controlled Trials

Léa Orsini^{1,2}, Andres Cardona², Emmanuel Lesaffre³, David Dejardin², Gwénaël Le Teuff¹

¹Oncostat U1018, Inserm, University Paris-Saclay, Villejuif, France

²Product Development, Data Sciences, F. Hoffmann-La Roche AG, Basel, Switzerland

³I-Biostat, KU-Leuven, Leuven, Belgium

Introduction The difference in restricted mean survival time between two arms (dRMST) is a meaningful measure of treatment effect in randomized controlled trials (RCTs) for time-to-event data, especially with non-proportional hazards. Choosing the time window $[0, \tau]$ is important to avoid any misinterpretation. Correct RMST estimation can be performed up to τ defined as the last follow-up time under a mild condition on the censoring distribution [1]. However, extensive comparisons between the different ways of selecting τ are still needed to address this important choice in practical settings. The objective is to empirically evaluate them through RCTs.

Methods Four techniques for choosing τ are evaluated: (a) 90th or 95th percentile of event times, (b) 90th or 95th percentile of follow-up times, (c) largest time with standard error of survival estimate within 5%, 7.5%, or 10%, and (d) minimum of the maximum follow-up times in each arm. τ -RMST estimations were performed using three frequentist methods (Kaplan-Meier estimator, pseudo-observations-based model, and Cox-based model) and two Bayesian methods (non-parametric model with a mixture of Dirichlet processes prior and pseudo-observations-based model), some of them allowing for covariate adjustments. For evaluation, we used three RCTs (IPSOS n=453, IMpower110 n=554, IMpower133 n=403) comparing immunotherapy with chemotherapy in lung cancer, with delayed treatment effects.

Results The range of τ calculated from the different techniques exceeded two years for IPSOS and IMpower110, and one year for IMpower133, impacting the Kaplan-Meier-based RMST estimation and its variance. With a delayed treatment effect, higher τ provides higher dRMST estimates with larger variances. Approaches (a) and (b) provide smaller τ often leading to immature conclusions while (d) results in an increased variability that can be mitigated in some cases by adjusting for appropriate covariates. Approach (c) emerged as a good candidate, balancing statistical precision with clinical relevance. All RMST estimators (frequentist and Bayesian) provided similar results.

Conclusion There is so far no consensus on defining τ , highlighting the need for clearer guidelines and greater transparency. Ideally, τ should be defined a priori with a clinical rationale. If not, data-driven approaches can be employed. Based on our findings, we recom-

mend the (c) proposal as it ensures sufficient representation of patients at risk. Establishing standardized, clinically relevant practices for defining τ will enhance the applicability and reproducibility of RMST analyses in future research.

[1] Lu Tian et al. On the Empirical Choice of the Time Window for Restricted Mean Survival Time (2020), *Biometrics*, 76(4): 1157–1166.

36: Identifying Risk Factors for Hospital Readmission in Home-Based Care: a Study from a Monographic Paediatric Cancer Centre

Sara Perez-Jaume^{1,2}, Maria Antònia Colomar-Riutort², Anna Felip-Badia¹, Maria Fabregat¹, Laura Andrés-Zallo¹

¹BiMaU, Sant Joan de Déu Pediatric Cancer Center Barcelona, Spain

²Department of Basic Clinical Practice, Universitat de Barcelona, Spain

Introduction Paediatric cancer is a group of rare malignancies that occur in childhood and adolescence. This potentially life-threatening disease often requires aggressive therapies, such as chemotherapy or immunotherapy. The nature of these interventions requires patients to be hospitalised multiple times. In this context, a monographic paediatric cancer centre in the south of Europe initiated a home-based hospitalisation programme for paediatric patients diagnosed with cancer, which potentially offers relevant benefits (enhanced quality of life and reduced economic costs). However, a concern with home-based hospitalisations is the occurrence of adverse events, such as the need for hospital readmission during the hospitalisation at home, which is considered an unfavourable outcome in home-based care. Data from this home hospitalisation programme are available from its foundation in November 2021 until June 2024. The aim of this work is to use these data to identify risk factors for hospital readmission during the home-based hospitalisation.

Methods The dataset used in this project poses a statistical challenge since patients may be hospitalised at home more than once. Appropriate methods for repeated measures are then required for a proper analysis. Since the outcome of interest is the binary variable "need for hospital readmission during the home hospitalisation", we used Generalized Estimating Equations (GEE) and Generalized Linear Mixed Models (GLMM) with a logit link function (marginal/subject-specific approaches). From these models, we derive the corresponding odds ratios. We applied a variable selection algorithm to identify risk factors for hospital readmission.

Results Data consist of the 380 home-based hospitalizations from 156 paediatric patients

previously diagnosed with cancer included in the home hospitalisation programme. Most patients were male (59%) and the median distance from hospital to the place of home-based hospitalisation was 8 km. Both GEE and GLMM approaches led to a final model with four variables; being three of them significantly associated with the outcome. Among the reasons for the home-based hospitalisation, we found that hydration-intended hospitalisations reduced the odds of hospital readmission compared to the rest of reasons considered. Moreover, lower neutrophil counts increased the odds of hospital readmission. The occurrence of incidences with the intravenous route also increased the odds of hospital readmission.

Conclusion We identified reason of hospitalisation, neutrophil count and the occurrence of incidences with the intravenous route as risk factors for hospital readmission in the context of home-based care in paediatric oncology, which might influence physicians' decisions about the management of these patients at home.

37: A Web-Application for Predicting Serious Adverse Event for Guiding the Enrollment Procedure in Clinical Trials with Machine Learning Methods

Ajsi Kanapari, Corrado Lanera, Dario Gregori

Unit of Biostatistics, Epidemiology and Public Health, University of Padova, Padova, Italy, Italy

Background Serious Adverse events (SAEs) refer to the undesired occurrence of an event that derives from a drug reaction with direct consequences on patients' life and compromising study validity and safety. On the matter there is room for improvement, that can be guided by the usage of Machine Learning (ML) with the aim of identifying subgroups of patients with meaningful combination of clinical features linked to SAEs, for limiting their frequency, with the usage of probabilistic methods that rely on clinical features rather than on specific dichotomized variables. However, they are explored often through post-hoc analysis and not directly informing the design of Clinical Trials, due to their complex application in a dynamic context, which makes necessary the support of electronic applications.

Objective The aim of this work is the development of a framework and a web-application in accordance with FDA guidance on enrichment strategies¹ for reducing trial variability, that implements ML models to allow early detection. It employs ML models to identify patients at high risk of SAEs, enhancing early detection and informing inclusion/exclusion decisions. Historical data of early phase trials are used to train predictive models that estimate SAE probabilities for new participants, with inclusion decisions guided by a predefined decision rule.

Results Simulations and the application on a case study assess the operational characteristics of the proposed framework, with the aim to maintain balance between the reduction of SAEs incidence, algorithm accuracy and maintaining generalizability of the study. Due to reduced variability consequent to patient exclusion, and most importantly the reduction of drop-outs lead to having power not only maintained but also increased, if the model provides a high performance, however issues are found particularly if low specificity is involved that would cause the unnecessary exclusion of low risk of SAEs subjects. On the positive extend, the algorithm provides reduced standard errors and more precise estimates of treatment effect.

38: Assessing the Overall Burden of Adverse Events in Clinical Trials: Approaches and Challenges

Seid Hamzic, Hans-Joachim Helms, Eva Rossman, Robert Walls

F. Hoffmann-La Roche Ltd, Basel, Switzerland

Measuring the total toxicity or adverse event (AE) burden of a therapeutic intervention is a longstanding challenge in clinical research. While trial reports commonly provide the incidence of individual AEs or a summary of the proportion of patients experiencing serious, e.g. grade 3 AEs, these metrics do not necessarily capture the global burden that they may impose on patients. Various approaches to consolidate AEs into a single composite score, such as summing CTCAE grades, have been proposed. However, these efforts face substantial methodological and interpretational hurdles.

This work offers a theoretical exploration of how AE burden could be conceptualized, quantified, and used when comparing two or more therapies. We review the limitations of incidence-based reporting that fails to capture interdependence or cumulative effects of multiple, possibly lower-grade AEs. We then discuss the existing proposals for composite toxicity scoring, noting the difficulties in weighting different AEs, some of which might be more tolerable to patients despite a higher grade. Additionally, current standard data collection approaches might lack the granularity necessary to distinguish differences in patient experience or quality of life.

We argue that while composite scores can offer a more holistic view of the total harm posed by a drug, they risk oversimplification and obscuring the clinical relevance of specific, important toxicities. Ultimately, this highlights a need for more robust data collection and careful methodological development that balances interpretability and accuracy in the comparison of AE burden across treatments.

Tuesday Posters at Biozentrum

Tuesday, 2025-08-26 09:15 - 10:45, Biozentrum, 2nd floor

1: Advantages and Pitfalls of a Multi-Centre Register Collecting Long-Term Real-World Data on Medical Devices: Insights from a Cochlear Implant Registry

Karin A. Koinig, Magdalena Breu, Jasmine Rinnofner, Stefano Morettini, Ilona Anderson

MED-EL Medical Electronics, Austria

Background There is a need for real world data (RWD) to demonstrate how medical devices function outside the setting of clinical studies and over longer time periods. One way to address this, is to establish registries collecting data from routine clinical visits. Here we present our experience from evaluating pre-surgery to two years post-surgery data from a multicentre cochlear implant registry.

Methods Data were extracted in anonymized form from a registry covering five clinics. The medical devices studied were cochlear implants, which help individuals with severe to profound sensorineural hearing loss (deafness) to regain their hearing. Key outcomes included speech perception, wearing time of the implant, self-perceived auditory benefit, self-reported quality of life, and safety results.

Results The registry provided extensive data but revealed differences in clinical practices, which made summarizing data across different assessments a challenge. Not all clinics collected the same information, although a minimal measurement data set was specified in the registry protocol. For example, the methods used to assess speech perception varied between centres, including differences in noise levels and test formats. In addition, we observed a high dropout rate, which represents a possible bias: Particularly at long-term follow-up visits, those with more problems seemed more likely to return to the clinic, while those with fewer problems were more likely to be adequately cared for by the outpatient clinics and therefore more likely to be lost to follow-up. Overall, this resulted in a substantial amount of missing data, which was difficult to explain to regulatory bodies like the FDA and TÜV. To address this issue, we presented demographics and outcomes with and without the patients lost to follow-up.

Conclusion RWD are valuable but pose a challenge when collected in routine clinical practice, as the diversity of assessments and tests leads to different reporting standards and data gaps

that make it difficult to obtain homogeneous and usable data. Statisticians must work with the study team to develop clear and transparent strategies for data collection and data extraction to achieve consistent and reliable results from registries.

2: Development and Validation of Prognostic Models in Phase I Oncology Clinical Trials

Maria Lee Alcober^{1,2}, Guillermo Villacampa¹, Klaus Langohr²

¹Statistics Unit, Vall d'Hebron Institute of Oncology (Spain)

²Department of Statistics and Operations Research, Universitat Politècnica de Catalunya (Spain)

Phase I trials are an essential part of the development of oncology research. For patients, balancing the potential risks of toxicity against the benefits of investigational drugs is crucial. Consequently, participation in phase I trials requires a minimum life expectancy and the absence of relevant symptoms. However, in clinical practice, no objective measures are used to evaluate these criteria, and decisions rely on subjective judgment. Considering this background, this study aims to use different statistical methods to develop and validate prognostic models to better identify oncology patients who may benefit from early-phase clinical trials.

A total of 921 patients treated at the Vall d'Hebron Institute of Oncology from January 2011 to November 2024 were included in this study (799 in the development cohort and 122 in the validation cohort). Different strategies were used to develop the prognostic models: i) stratified Cox's proportional hazards models, ii) stratified Cox models enhanced with restricted cubic splines to address non-linearity, and iii) machine learning techniques such as decision trees and random survival forests to capture complex interactions. Risk scores derived from these models provide interpretable summaries of patient risk profiles, facilitating practical clinical use.

Results were validated using i) internal validation employing bootstrapping and cross-validation and ii) external validation using an independent dataset. Model performance was evaluated through discrimination (C-statistic), calibration (calibration plots and the Hosmer-Lemeshow test), overall performance (Brier score), and clinical utility (decision curve analysis).

Internal validation consistently outperformed external validation across all performance metrics, particularly in calibration and clinical utility. Among the models, random survival forests achieved the highest C-statistic, demonstrating superior discrimination. Conversely, incor-

porating restricted cubic splines into the Cox's proportional hazards model did not notably improve the evaluated metrics.

This work offers a replicable framework for deriving and validating risk scores that improve precision in patient selection for phase I trials. Future efforts will focus on formalising calibration methods and comparing these models and scores with other published prognostic tools using external validation.

3: Application of Bayesian Surrogacy Models to Select Primary Endpoint in Phase 2 Based on Relationship to a Phase 3 Endpoint

Alexandra Jauhainen¹, Enti Spata², Patrick Darken³, Carla A. Da Silva⁴

¹R&I Biometrics and Statistical Innovation, Late Respiratory & Immunology, BioPharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden

²R&I Biometrics and Statistical Innovation, Late Respiratory & Immunology, BioPharmaceuticals R&D, AstraZeneca, Cambridge, UK

³R&I Biometrics and Statistical Innovation, Late Respiratory & Immunology, BioPharmaceuticals R&D, AstraZeneca, Gaithersburg, US

⁴Early Respiratory and Immunology Clinical Development, BioPharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden

Background A key goal of treatment in asthma is to prevent episodes of severe symptom worsening called exacerbations. Designing trials for these relatively rare events is a challenge, especially in the early phases of development of new therapies, as the studies tend to be large and lengthy. Hence, exacerbations are not usually studied as a primary endpoint until phase 3. Alternative endpoints to use in early phase trials of shorter duration can be lung function measurements like FEV₁, or the novel endpoint CompEx, which is a composite endpoint enriching exacerbations by adding events defined from deteriorations in diary card variables.

Methods All three endpoints; FEV₁, CompEx, and exacerbations; were evaluated using patient level data across a set of 14 trials with 27 treatment comparisons. FEV₁ was analysed as change from baseline, while CompEx and exacerbations were modelled both in a time-to-first and recurrent event setting, across two timeframes (3- and 12-months duration).

Bayesian bivariate random-effect meta-analysis was applied to estimate the total correlation for treatment effects for FEV₁ and CompEx with exacerbations. Bayesian surrogacy analysis within the Daniels & Hughes framework was applied across treatment comparisons to evaluate

the trial-level relationship between CompEx and exacerbations.

Results The change from baseline in FEV₁ at 3 months had a weak correlation with the preferred phase 3 endpoint, the rate ratio for exacerbations at 12 months, and showed limitations in its ability to quantify the effect reported on exacerbations across drug modalities.

In contrast, the CompEx hazard ratio at 3 months correlated well with the 12-month rate ratio observed on exacerbations. CompEx was confirmed as a surrogate in terms of predicting treatment effects observed on exacerbations, with a high level of correspondence between the endpoints across modalities and asthma severities.

Conclusion FEV1 remains an important respiratory endpoint, especially for drugs with bronchodilating properties, but has limitations as a primary phase 2 endpoint across modalities when the aim is to target exacerbations in phase 3.

CompEx has an increased event frequency compared to exacerbations alone, especially noticeable in populations with low exacerbation rates (mild/moderate asthma). This makes CompEx an attractive endpoint to use in design of early phase trials across a range of modalities, especially towards the milder spectrum of disease, substantially reducing sample sizes needed.

This research was funded by AstraZeneca.

4: Discontinuation and Attrition Rates in Phase II or Phase III First-Line Randomized Clinical Trials (RCTs) of Solid Tumors

Virginia Delucchi¹, Chiara Molinelli², Luca Arecco², Andrea Boutros³, Davide Soldato², Matteo Lambertini^{2,3}, Dario Trapani^{4,5}, Bishal Gyawali⁶, Gabe S Sonke⁷, Sarah R Brown⁸, Matthew R Sydes^{9,10}, Luca Boni¹, Saskia Litiere¹¹, Eva Blondeaux¹

¹U.O. Epidemiologia Clinica, IRCCS Ospedale Policlinico San Martino, Genova, Italy

²U.O.C. Clinica di Oncologia Medica, IRCCS Ospedale Policlinico San Martino, Genova, Italy

³Department of Internal Medicine and Medical Specialties, University of Genova, Genova, Italy

⁴Division of New Drugs and Early Drug Development for Innovative Therapies, European Institute of Oncology, IRCCS, Milan 20141, Italy

⁵Department of Oncology and Hemato-Oncology, University of Milan, Milan 20122, Italy

⁶Division of Cancer Care and Epidemiology, Cancer Research Institute, Queen's University,

Kingston, ON, Canada

⁷Division of Medical Oncology, Netherlands Cancer Institute, Amsterdam, the Netherlands

⁸Leeds Cancer Research UK Clinical Trials Unit, University of Leeds, Leeds, UK

⁹BHF Data Science Centre, HDR UK, London, UK

¹⁰Data for R&D, Transformation Directorate, NHS England, London, UK

¹¹EORTC Headquarters, Brussels, Belgium

Background

Differential discontinuation and attrition rates in randomized controlled trials (RCTs) bias efficacy assessments, potentially leading to misinterpretations of treatment effects. Despite their critical role, the extent and implications of these rates in cancer trials remain unclear. We aimed to systematically quantify discontinuation and attrition rates in RCTs of solid tumors and how variation in these rates might impact the estimated treatment effect on overall survival.

Methods A systematic review of published literature was carried out to identify phase II or phase III RCTs of first-line treatments published from Jan-2015 to Feb-2024 of solid tumors in Medline. Reported treatment discontinuation and post-study treatments figures were extracted from CONSORT diagram and/or text. Attrition was computed as the percentage of patients reported as discontinuing study drugs for whom a post-study treatment was not documented. We investigated differences in discontinuation and attrition rates according to type of cancer, sponsor and trial phase. Discontinuation and attrition by treatment arm were not reported due to potential influence of experimental treatment on progression. Simulations evaluating the impact of different discontinuation and attrition rates on overall survival will be implemented and presented at the congress.

Results Out of 22,141 records screened, 533 trials met the inclusion criteria. The majority (56%) were phase III, industry-sponsored (54%) trials; 126 (24%) trials enrolled patients with non-small cell lung cancer, 79 (15%) breast cancer, 53 (10%) colorectal cancer, 40 (8%) other gastrointestinal cancers, 29 (5%) melanoma, 28 (5%) pancreatic cancer and 178 other tumor types. Treatment discontinuation figures were reported in 415 (78%) trials, with a median patient discontinuation rate of 83%. No difference in the patient' treatment discontinuation rate was observed according to sponsor and trial phase. Among the 415 trials reporting patient' treatment discontinuation, data on any post-study treatment was reported in 220 (53%) trials. Median patient attrition rate was 37%. The highest median patient attrition rate was observed for urothelial cancer trials (53%) and the lowest for breast cancer trials (28%). Industry-sponsored trials reported a higher median patient attrition rate than academic trials (38% vs 26%, respectively). No difference in patient attrition rate was observed between phase II and phase III trials.

Conclusions Although most cancer trials published on treatment discontinuation rates, post-study treatments were less frequently documented. Our results highlight the need to improve the reporting of these figures to ensure transparency, reliability, and accurate assessment of treatment effects on long-term outcome measures.

5: Enhancing Dose Selection in Phase I Cancer Trials: Extending the Bayesian Logistic Regression Model with Non-DLT Adverse Events Integration

Luca Genetti, Andrea Nizzardo, Marco Pergher

Evotec - Verona, Italy

This work presents the Burdened Bayesian Logistic Regression Model (BBLRM), an enhancement to the Bayesian Logistic Regression Model (BLRM) for dose-finding in phase I oncology trials. Traditionally, the BLMR determines the maximum tolerated dose (MTD) based on dose-limiting toxicities (DLTs)¹. However, clinicians often perceive model-based designs like BLMR as complex and less conservative than rule-based designs, such as the widely used 3+3 method^{2,3}. To address these concerns, the BBLRM incorporates non-DLT adverse events (nDLTAEs) into the model. These events, although not severe enough to qualify as DLTs, provide additional information suggesting that higher doses might result in DLTs.

In the BBLRM, an additional parameter δ is introduced to account for nDLTAEs. This parameter adjusts the toxicity probability estimates, making the model more conservative in dose escalation without compromising the accuracy in allocating the true MTD. The δ parameter is derived from the proportion of patients experiencing nDLTAEs and is tuned based on the design characteristics to balance the model's conservatism. This approach aims to reduce the likelihood of assigning toxic doses as MTD while involving clinicians more directly in the decision-making process identifying the nDLTAEs along the study conduction.

The work includes a simulation study comparing BBLRM with more traditional versions of BLMR^{4,5} and a two stage Continual Reassessment Method (CRM)⁶ that incorporates nDLTAEs across various scenarios. The simulations demonstrate that BBLRM significantly reduces the selection of toxic doses as MTD without compromising the accuracy of MTD identification. These results suggest that integrating nDLTAEs into the dose-finding process can enhance the safety and acceptance of model-based designs in phase I oncology trials.

References 1. Neuenschwander B et al. Critical aspects of the bayesian approach to phase I cancer trials. ***Statistics in Medicine*** 2008.

2. Love SB et al. Embracing model-based designs for dose-finding trials. *British Journal of Cancer* 2017.
3. Kurzrock R et al. Moving beyond 3+3: the future of clinical trial design. *American Society of Clinical Oncology Educational Book* 2021.
4. Zhang H et al. Improving the performance of Bayesian logistic regression model with overdose control in oncology dose-finding studies. *Statistics in Medicine* 2022.
5. Ghosh D et al. Hybrid continuous reassessment method with overdose control for safer dose escalation. *Journal of Biopharmaceutical Statistics* 2023.
6. Iasonos A et al. Incorporating lower grade toxicity information into dose finding designs. *Clinical Trials* 2011.

6: Bayesian Inference of the Parametric Piecewise Accelerated Failure Time Models for Immune-Oncology Clinical Trials

Xingzhi Xu, Satoshi Hattori

Osaka University, Japan

Modeling delayed treatment effects pose significant challenges in survival analysis, particularly in immune-oncology trials where Kaplan-Meier curves often exhibit overlapping patterns. Overlapping Kaplan-Meier curves implies the proportional hazard assumption is violated and the use of hazard ratio to summarize treatment effects is not appealing. In addition, it implies some patients do not benefit from the immuno-oncology drug. To address these issues, Sunami and Hattori (2024) introduced the piecewise Accelerated Failure Time (pAFT) model, employing a frequentist semi-parametric maximum-likelihood approach to account for delayed treatment effects and to evaluate each patient's probability of receiving benefit from the treatment. Their framework, while innovative, faced challenges in handling complex treatment-by-covariates interactions.

Building on their foundational work, this paper introduces two Bayesian parametric extensions: the pAFT model and the interactive piecewise Accelerated Failure Time (ipAFT) model. The Bayesian framework enhances the original model by incorporating prior knowledge and improving parameter estimation precision. The ipAFT model, in particular, extends the methodology by explicitly modeling treatment-by-covariates interactions, offering deeper insights into treatment efficacy on different subgroups.

Comprehensive simulation studies demonstrate that the proposed Bayesian models perform exceptionally in capturing delayed treatment effects, achieving accurate estimations and reliable coverage probabilities even with small sample sizes. The ipAFT model provides two measures for patient-specific treatment effects: probabilities of receiving the benefit from the treatment and patient-specific benefit after the delayed time. Applying some multivariate analysis techniques (such as hierarchical clustering) to the two measures, we can effectively characterize patients' treatment effects. Application to a real-world immuno-oncology clinical trial dataset reveals distinct patient subgroups based on the result of the ipAFT model.

By addressing key limitations of traditional survival models and extending Sunami and Hattori's pAFT framework, the proposed Bayesian models offer flexible tools for analyzing immuno-oncology clinical trials. The stable and flexible natures allow our methods to be useful in early-phase clinical trials with small patient counts.

7: Bayesian Power-Based Sample Size Determination for Single-Arm Clinical Trials with Time-to-Event Endpoints

Go Horiguchi¹, Isao Yokota², Satoshi Teramukai¹

¹Department of Biostatistics, Graduate School of Medical Science, Kyoto Prefectural University of Medicine, Japan

²Department of Biostatistics, Hokkaido University Graduate School of Medicine, Japan

Introduction Single-arm exploratory trials are widely used in early-phase oncology research to assess the potential of new treatments, often using time-to-event endpoints. Conventional sample size calculations under a frequentist framework typically rely on limited statistics, such as point estimates of survival rates at specific time points or a single hazard ratio (HR). By contrast, Bayesian methods can incorporate prior information and allow interim decisions with greater flexibility. We propose a Bayesian sample size determination method based on posterior and prior predictive probabilities of the hazard ratio, introducing analysis and design priors to improve decision-making accuracy and efficiency.

Methods In our Bayesian design, we set a target hazard ratio of 1 to show the superiority of new treatment. Using the analysis prior, we compute the posterior probability that the hazard ratio is below this target. If this probability exceeds a prespecified threshold, we conclude efficacy and stop the trial. For each candidate sample size, we draw from the design prior, generate predicted outcomes under proportional hazards, and calculate the proportion of simulated trials that would meet the stopping criterion. This proportion is the Bayesian power. The smallest sample size achieving the desired power is then selected. Here,

the analysis prior encodes historical knowledge about the parameter, while the design prior represents its uncertainty at the planning stage.

Results Simulation results show that more informative analysis priors reduce sample size, while greater uncertainty in the design priors increases it. For designs without interim analysis, the Bayesian method produces sample sizes comparable to or smaller than frequentist methods while maintaining type I error rates. Interim analyses reduce expected sample size and trial duration, with thresholds for posterior probabilities influencing early termination probabilities. Results also demonstrate flexibility in accommodating varying assumptions about survival distributions and parameter uncertainties.

Conclusion The proposed Bayesian sample size determination method efficiently incorporates prior information and interim analyses, making it a practical alternative to traditional frequentist approaches. This approach enables flexible and rational trial designs, reducing conflicting decisions and improving resource use. Limitations include reliance on the proportional hazards assumption and computational demands for simulation-based power calculations. Future research should explore extensions to handle censoring and other complexities in clinical trials.

8: Calibration of Dose-Agnostic Priors for Bayesian Dose-Finding Trial Designs with Multiple Outcomes

Emily Alger¹, Shing M. Lee², Ying Kuen K. Cheung², Christina Yap¹

¹The Institute of Cancer Research, United Kingdom

²Columbia University, USA

Introduction The goal of dose-finding oncology trials is to assess the safety of novel anti-cancer treatments across multiple doses and to recommend dose(s) for subsequent trials. Based on previous observed responses, trialists dynamically recommend new doses for further investigation during the trial.

Adaptive decision making lends itself to Bayesian learning, with Bayesian frameworks increasingly guiding dose recommendations in model-based dose-finding designs, such as the Continual Reassessment Method (CRM) design. However, these approaches often add complexity by incorporating multiple outcomes and require appropriate prior selection. For trialists who lack prior knowledge, we may look to adopt a dose-agnostic prior – with each dose equally likely to be the a priori optimal dose. However, applying existing methodology to a multiple-outcome CRM may inflate suboptimal, low dose recommendations.

Methods We broaden calibration techniques for single-outcome trial designs to calibrate dose-agnostic priors for multiple-outcome trial designs, such as designs that jointly evaluate Dose Limiting Toxicities (DLTs) and efficacy responses, or DLTs and patient-reported outcomes (PROs). The a priori probability each dose is identified as the recommended dose is written analytically and optimised using divergence minimisation. A simulation study is presented to demonstrate the effectiveness of calibrated priors for both the PRO-CRM[1] trial design and the joint-outcome CRM model proposed by Wages and Tait[2] in comparison to marginally calibrated priors.

Results Our analytical and computationally efficient technique maintains an a priori dose agnostic prior whilst improving the probability of correct selection (PCS) and standard deviation of PCS across most simulation scenarios. Thus, jointly calibrated priors reduce the bias present in simulation performance with marginally calibrated priors.

Conclusion Leveraging analytical expressions for a priori optimal dose recommendations enables computationally efficient implementation and reduces the need for extensive simulations to confirm trial design performance. What's more, this approach supports trialists to develop deeper intuition about their prior choices, thus strengthening their confidence in selecting robust and suitable priors. As Bayesian dose-finding trial designs continue to advance, research and guidance on the effective calibration of design parameters is essential to support the uptake of Bayesian designs, demonstrate the importance of rigorous prior calibration, and ensure optimal performance in practice.

[1] Lee, Shing M., Xiaoqi Lu, and Bin Cheng. "Incorporating patient-reported outcomes in dose-finding clinical trials." *Statistics in medicine* 39.3 (2020): 310-325.

[2] Wages, Nolan A., and Christopher Tait. "Seamless phase I/II adaptive design for oncology trials of molecularly targeted agents." *Journal of biopharmaceutical statistics* 25.5 (2015): 903-920.

9: Estimands in Platform Trials with Time - Treatment Interactions

Ziyan Wang, Dave Woods

Statistical Sciences Research Institute (S3RI), University of Southampton, United Kingdom

Background

In long-running platform trials, treatment effects may change over time due to shifts in the recruited population or changes in treatment efficacy—such as increased clinician experience

with a novel surgical technique [1]. Most existing studies have assumed equal time trends across treatment arms and controls, focusing on treatment-independent time effects [2,3]. However, when time trends are unequal between treatment arms and controls, the standard estimands can lead to inflated type I error rates, reduced statistical power, and biased treatment effect estimates. In this study, we propose a novel model-based estimand designed to correct for unequal time trends, thereby ensuring robust and accurate inference in platform trials.

Methods

We propose a general model-based estimand based on a time-averaged treatment effect that is adaptable to a variety of time trend patterns in platform trials. In our study, we compare the performance of the standard treatment effect estimand with our generalized estimand in settings where time trends differ between treatment arms and the control. A simulation study is conducted within the framework of Bayesian platform trials—including those employing response-adaptive randomization (RAR)—and performance is evaluated in terms of error rates, bias, and root mean squared error.

Results

Our findings demonstrate that the generalized estimand is robust across various time trend patterns, including nonlinear trends. Flexible modelling with this estimand maintains unbiasedness and reduces power loss compared to the standard estimand. Moreover, the approach remains effective under adaptive randomization rules. All simulation analyses were performed using our “*BayesianPlatformDesignTimeTrend*” R package, which is publicly available on CRAN.

Conclusion

This work provides a practical and innovative approach for addressing time trend effects in platform trials, offering new insights into the analysis of trials where unequal strength of time trends exists.

[1] K. M. Lee, L. C. Brown, T. Jaki, N. Stallard, and J. Wason. Statistical consideration when adding new arms to ongoing clinical trials: the potentials and the caveats. *Trials*, 22:1–10, 2021.

[2] Roig, M. B., Krotka, P., Burman, C.-F., Glimm, E., Gold, S. M., Hees, K., Jacko, P., Koenig, F., Magirr, D., Mesenbrink, P., et al. (2022). On model-based time trend adjustments in platform trials with non-concurrent controls. *BMC medical research methodology*. 22.1, pp. 1–16.

[3] Marschner, I. C., & Schou, I. M. (2024). Analysis of Nonconcurrent Controls in Adaptive Platform Trials: Separating Randomized and Nonrandomized Information. *Biometrical Journal*, 66(6), e202300334.

10: A Graphical Approach to Subpopulation Testing in Biomarker-Driven Clinical Trial Design

Boaz Natan Adler¹, Valeria Mazzanti², Pantelis Vlachos², Laurent Spiess²

¹Cytel Inc., United States of America

²Cytel Inc., Geneva, Switzerland

Introduction As targeted therapies in Oncology are fast-becoming commonplace, clinical studies are increasingly focused on biomarker-driven hypotheses. This type of research, in turn, requires methods for subpopulation analysis and multiplicity comparison procedures (MCPs) for sound clinical trials. In our case study, we employed advanced statistical software to design and optimize such a clinical study with a novel graphical approach to testing sequence and procedures.

Methods For this optimization exercise, we interrogated the typical areas of design interest: selecting an appropriate sample size, required number of events, and the timing and attributes of an interim analysis for the study. In addition, our optimization aim included a focus on the testing sequence of the study's subpopulations, biomarker-positive, and -negative, as well as a test of the overall study population. We also sought to optimize the MCP employed for the study, examining logrank and stepdown logrank tests, alongside different options for alpha splitting among the tests. Design variations and simulation were conducted using advanced statistical software and relied on a graphical approach to testing sequence and alpha splitting, in addition to visualizations of other study parameter variations.

Results This extensive simulation and optimization work allowed us to select a design that was tailored to the unique treatment effect assumptions of the investigational drug. We were able to convey design tradeoffs and the implications of testing sequence selection and other key design parameters in a graphical, relatable manner to the entire drug development team.

Conclusion A graphical approach to designing complex subpopulation analysis-driven clinical trials enables biostatisticians to assess design tradeoffs and selections clearly, while easing design and simulation work, and enhancing communication with governance committees.

11: Optimizing Biomarker-Based Enrichment Strategies in Clinical Trials

Djuly Asumpta PIERRE Paul¹, Irina Irincheeva², Hong Sun³

¹Nantes University (France), Bristol-Myers Squibb (Switzerland)

²Bristol-Myers Squibb Boudry (Switzerland)

³Bristol-Myers Squibb Boudry (Switzerland)

Background Identifying patients' groups based on biomarkers is crucial in oncology. Validating a biomarker as a stratification criterion in clinical trials can take several years. Choosing the threshold for continuous biomarkers is particularly challenging, often relying on a limited number of values evaluated with simplistic statistical approaches. Early dichotomization ignores the actual distribution of values and the potentially informative "grey zone".

Methods In this work, we adapt a biomarker enrichment design to identify the optimal threshold to determine patients who will benefit the most from the experimental treatment. We simulate Simon & Simon design for binomial endpoint and survival endpoint. Various scenarios of chosen thresholds are studied through simulations inspired by existing studies. ROC curve-based approach to determine the threshold, as well as the Song-Chi closed test procedure to assess the treatment effect in both the overall population and the biomarker-positive subgroups are explored.

Results Initial results suggest that our proposal effectively controls the Type I error for both binomial and survival endpoints. Additionally, switching to a ROC-curve approach for estimating the biomarker threshold improves statistical power by approximately 14%. Furthermore, incorporating the Song-Chi method allows testing of the difference in treatment effects between the standard control group and the experimental group in both the overall population (all the patients enrolled in the trial) and among the biomarker-positive patients, the patients most likely to benefit from the treatment. This method maintains rigorous Type I error control while still ensuring adequate power. Moreover, it facilitates the detection of treatment-specific fluctuations and subgroup dynamics within these two populations, leading to a more nuanced and precise analysis.

Conclusion In conclusion, this study highlights the importance of a more nuanced approach in selecting biomarker thresholds and improving biomarker enrichment strategy for clinical trials, which is essential to accelerate the development of personalized therapies while optimizing the efficiency of clinical trials.

Reference Simon N, Simon R. Adaptive enrichment designs for clinical trials.

Biostatistics. 2013 Sep;14(4):613-25. doi: 10.1093/biostatistics/kxt010.

Epub 2013 Mar 21. PMID: 23525452; PMCID: PMC3769998.

Song Y, Chi GY. A method for testing a prespecified subgroup in clinical trials

Stat Med. 2007 Aug 30;26(19):3535-49.

doi: 10.1002/sim.2825.PMID: 17266164

12: Leveraging Synthetic Data for Enhanced Clinical Research Outcomes

Szymon Musik^{1,2}, Agnieszka Kowalewska³, Gianmarco Gallone³, Jacek Zalewski³, Joanna Sasin-Kurowska³

¹Late Phase Global Clinical Data Management, Clinical Data & Insights, BioPharmaceutical Clinical Operations, R&D, AstraZeneca, Warsaw, Poland

²Department of Education and Research in Health Sciences, Medical University of Warsaw, Poland

³Clinical Programming, Clinical Data & Insights, BioPharmaceutical Clinical Operations, R&D, AstraZeneca, Warsaw, Poland

Background / Introduction In recent years, the pharmaceutical industry has been under immense pressure to make drug development faster and more efficient. Traditional clinical trials often face obstacles like high costs, prolonged durations, and challenges in participant recruitment, particularly for rare diseases. Additionally, testing of programming tools, databases, and software before acquiring patient data is cumbersome. Synthetic Data in Clinical Trials (SDCT) offers an innovative solution by providing high-quality, clinically realistic datasets that meet strict privacy conditions, facilitating thorough research.

Methods We developed AstraZeneca's Study Synthetic Data Tool (SYNCDATA), which generates synthetic data for a study (referred to as the target study) using its Architect Loader Spreadsheet (ALS) and data from an ongoing or completed study (referred to as the base study). Importantly, the target study may not yet have any data collected. Our pipeline leverages the event chronology specified by the ALS, allowing scenarios for each patient to be created before data generation. We categorize dataset variables into groups based on types, such as dates or binary options (e.g., Yes/No), and use designated methods for generating these variables. This approach employs classic statistical techniques like kernel density estimation and Bayesian networks. Designed primarily for study set-up testing, SYNCDATA explores potential variable values in the target study while preserving relationships from the base study. It can also incorporate incorrect values into the data if necessary.

Results Incorporating synthetic data into clinical trials has significantly improved data scarcity challenges. SYNCDATA generates synthetic data as soon as the ALS for a study is available, enabling users to test programming tools, databases, software, and visualizations. Further-

more, synthetic data supports data science projects. SYNTDATA is secure and ensures patient privacy.

Conclusion Synthetic data is set to transform clinical trials by addressing the current challenges in the pharmaceutical industry. It reduces development timelines and enhances data integration efficiency, allowing more reliable trial simulations. Adopting synthetic data as a vital component of clinical research could reshape conventional practices and usher in a new era of data-driven drug development.

13: Graph-Based Integration of Heterogeneous Biological Data for Precision Medicine: A Comparative Analysis of Neo4j and MySQL

Byoung Ha Yoon

KRIBB(korea research institute of bioscience and biotechnology), Korea, Republic of (South Korea)

Precision medicine aims to provide personalized treatment plans tailored to individual patients. However, the complexity and scale of biomedical data, coupled with the exponential growth of clinical knowledge derived from diverse biological databases and scientific publications, pose significant challenges in clinical applications. A key challenge in this context is understanding and integrating the intricate relationships between heterogeneous biological data types.

In this study, we address this challenge by integrating multiple biological datasets—such as protein-protein interactions, drug-target associations, and gene-disease relationships—into a unified graph database. The constructed graph consists of approximately 150,000 nodes and 100 million relationships, with data pre-processed to remove redundancies. To assess the suitability of graph-based databases for handling complex biological networks, we compared the performance of Neo4j, a state-of-the-art graph database, with MySQL, a traditional relational database. Our results demonstrate that while MySQL struggled with complex queries involving multiple joins, Neo4j exhibited superior performance, providing rapid responses to the same queries.

These findings emphasize the potential of graph databases for efficiently storing and querying complex biological relationships. Moreover, the interconnected nature of biological data in graph structures facilitates the application of computational biology techniques, such as network analysis and clinical biostatistics, to uncover hidden patterns and infer new insights. This approach not only enhances the understanding of biological systems but also holds promise

for improving clinical decision-making and advancing the field of precision medicine.

14: Revolutionizing Clinical Data Management: A Strategic Roadmap for Integrating AI/ML into CDM

Joanna Magdalena Sasin-Kurowska¹, Szymon Musik¹, Mariusz Panczyk²

¹Astra Zeneca, Poland

²Medical University of Warsaw, Poland

Clinical Data Management (CDM) is essential in clinical research, ensuring the accuracy and integrity of data for regulatory submissions. As clinical trials become more complex and generate larger volumes of data—especially in Phase III trials—there is a growing need for advanced tools to manage and analyze this information. This poster highlights key findings from our research on integrating Artificial Intelligence (AI) and Machine Learning (ML) into CDM, transforming it into Clinical Data Science (CDS). By reviewing literature from 2008 to 2025, we identified emerging trends such as the use of Natural Language Processing (NLP) to analyze unstructured data, AI/ML for automating data cleaning and analysis, and new technologies like blockchain, wearable devices, and patient-centric approaches. Our results indicate that AI/ML can improve data quality, automate processes, and enhance predictive analytics, offering a more efficient and scalable solution for clinical research. We also present a roadmap for successfully integrating AI/ML into CDM to drive innovation and advance clinical research. This review emphasizes the need for a strategic, multidisciplinary approach to fully leverage these technologies for more efficient and accurate clinical trials.

15: Strategies to Scale Up Model Selection for Analysis of Proteomic Datasets using Multiple Linear Mixed-Effect Models

Ilya Potapov, Matthew Davis, Adam Boxall, Francesco Tuveri, George Ward, Simone Jueliger, Harpreet Saini

Astex Pharmaceuticals, United Kingdom

Linear mixed-effect models (LMM) are a key tool to model biomedical data with dependencies. For example, longitudinal read-outs from patients would necessarily need to address the correlation between samples, which violates the assumption of independence in the standard

linear modelling approach. Designing the LMEM in terms of factors and their interaction that constitute the model is an elaborate process that takes into account both the formal analysis of the model variance and the end points of the study. Whereas there are multiple hypotheses of how best to design LMEMs, this process takes place normally at the level of a single model. In biomedical applications, however, we are often interested in multiple comparisons. In this case, the LMEM design process should be scaled up to optimise the model design for all comparisons simultaneously. In this work, we considered an example of the multiple design problem in a proteomic experiment. We showed how a general framework for the multiple LMEM designs can be established via the analysis of variance of the full and restricted (nested) models. This analysis included the formation of the P-value distribution for each of the factor terms and subsequent analysis of that distribution. We also demonstrated that the multiple design framework necessarily poses a question of whether all the models should have the same universal model design or individualised tailored models per protein. Both pathways are possible from the methodological point of view, yet they may have different implications for statistical inference. We discuss these implications.

16: Cost-Utility Analysis of Sodium-Glucose Cotransporter-2 Inhibitors on Chronic Kidney Disease Progression in Diabetes Patients: a Real-World Data in Thailand

Sukanya Siriyotha¹, Amarit Tansawet², Oraluck Pattanaprateep¹, Tanawan Kongmalai³, Panu Looareesuwan¹, Junwei Yang¹, Suparee Wisawapipat Boonmanunt¹, Gareth J McKay⁴, John Attia⁵, Ammarin Thakkinstian¹

¹Department of Clinical Epidemiology and Biostatistics, Faculty of Medicine Ramathibodi Hospital, Mahidol University, Bangkok, Thailand

²Department of Research and Medical Innovation, Faculty of Medicine Vajira Hospital, Navamindradhiraj University, Bangkok, Thailand

³Division of Endocrinology and Metabolism, Faculty of medicine Siriraj Hospital Mahidol University, Bangkok, Thailand

⁴Centre for Public Health, School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast, Belfast, United Kingdom

⁵School of Medicine and Public Health, and Hunter Medical Research Institute, University of Newcastle, New Lambton, New South Wales, Australia

Introduction and Objective(s): Type 2 diabetes (T2D) increases the risk of micro- and macro-vascular complications, including chronic kidney disease (CKD), a major burdens that could significantly impair the quality of life and socioeconomic status. Evidence from numerous clinical trials demonstrate the benefits of sodium-glucose co-transporter 2 inhibitors

(SGLT2is) in CKD prevention. However, the high cost of SGLT2i may limit their accessibility, despite economic evaluations suggesting cost-effectiveness. Therefore, this study aims to conduct a cost-utility analysis using real-world data in Thailand to provide more realistic and relevant evidence for policy decisions.

Method(s) and Results: Clinical and cost data of CKD patients between 2012 and 2022 were retrieved from Ramathibodi T2D data warehouse. Markov model was constructed for the following states: CKD stage 3, 4, 5, and death. A cost-utility analysis that estimates the cost per quality-adjusted life year (QALY) between two interventions: non-SGLT2i versus SGLT2i was performed in societal perspective. The incremental cost-effectiveness ratio (ICER) was calculated by dividing the difference in costs between the compared treatments by the difference in QALY associated with each treatment. A total of 20,735 patients were recruited. The lifetime costs were US\$72,234.98 and 74,887.31 in patients with renal replacement therapy (RRT) and US\$71,638.41 and 74,749.86 in patients without RRT, for non-SGLT2i and SGLT2i, respectively. ICERs were US\$955.40 and 1,114.56 per QALY in patients with and without RRT.

Conclusions SGLT2i was associated with higher treatment cost compared with non-SGLT2i. However, SGLT2i was still cost-effective considering Thailand willingness to pay at US\$4,651 per QALY.

Keywords Cost-utility analysis (QALY), Real-world data, Type 2 diabetes (T2D), Chronic kidney disease (CKD), Sodium-glucose co-transporter 2 inhibitors (SGLT2is)

References [1] Beckman JA, Creager MA. Vascular Complications of Diabetes. *Circulation Research*. 2016;118(11):1771-85.

[2] Wanner C, Inzucchi SE, Lachin JM, Fitchett D, von Eynatten M, Mattheus M, et al. Empagliflozin and Progression of Kidney Disease in Type 2 Diabetes. *N Engl J Med*. 2016;375(4):323-34.

[3] Reifsnyder OS, Kansal AR, Wanner C, Pfarr E, Koitka-Weber A, Brand SB, et al. Cost-Effectiveness of Empagliflozin in Patients With Diabetic Kidney Disease in the United States:

17: Comparing the Safety and Effectiveness of Covid-19 Vaccines Administered in England using OpenSAFELY: A Common Analytic Protocol

Martina Pesce¹, Christopher Wood¹, Helen McDonald², Frederica Longfoot¹, Venexia Walker³, Edward PK Parker⁴, William J Hulme¹

¹Bennett Institute for Applied Data Science, Nuffield Department of Primary Care Health Science, Oxford University, UK

²University of Bath, UK

³Population Health Sciences, Bristol Medical School, University of Bristol, UK

⁴London School of Hygiene and Tropical Medicine, UK

Background In England, Covid-19 vaccination campaigns have been delivered in Spring and Autumn each year since 2021, and this pattern is set to continue for the foreseeable future. At least two vaccine products are used each campaign to mitigate any potential unforeseen supply or safety issues.

Post-authorisation evaluations of these vaccines in routine, out-of-trial settings are crucial: incidence of longer-term and rarer outcomes are often not reliably estimable in trials, and vaccines may perform differently in more diverse population groups or in the context of newer viral variants.

The regularity and similarity of campaigns, including future campaigns, coupled with the availability of reliable routinely-collected health data on who is getting which vaccine and when, provides an opportunity to specify a single analysis protocol that can be reused across multiple campaigns.

Methods We developed a Common Analytic Protocol to compare the safety and effectiveness of vaccine products used in each Covid-19 vaccination campaign. Planned analyses will use the OpenSAFELY research platform which provides secure access to routinely-collected health records for millions of people in England.

The protocol uses complementary approaches to control for confounding (one-to-one matching without replacement and inverse probability of treatment weighting) to compare products for a variety of safety and effectiveness endpoints, within a variety of population subgroups, and with various accompanying sensitivity analyses and balance checks. The analogous hypothetical randomised trial that the design emulates is also described.

All design elements are specified explicitly in R scripts, fully executable against simulated dummy data before any real data is available for analysis.

Discussion The ability to plan analyses comparing vaccine products well in advance of the delivery of the campaign has numerous benefits and challenges, which will be described in this talk. We invite feedback on the proposed design prior to its use in real data.

18: Statistical Requirements in Medical Diagnostic Development Across the UK, US, and EU Markets: A Review of Regulation, Guidelines and Standards.

Timothy Hicks^{1,2}, Joseph Bulmer¹, Alison Bray¹, Jordan L. Oakley³, Rachel L. Binks^{2,3}, Kile Green², Will S. Jones⁴, James M.S. Wason³, Kevin J. Wilson³

¹Newcastle Upon Tyne Hospitals NHS Foundation Trust, United Kingdom

²NIHR HealthTech Research Centre in Diagnostic and Technology Evaluation, United Kingdom

³Newcastle University, United Kingdom

⁴Centre of Excellence for Data Science, Artificial Intelligence and Modelling (DAIM), University of Hull, United Kingdom

Background When developing novel medical diagnostic devices, including In Vitro Diagnostics, Medical Diagnostic Software, and General Medical Devices, developers must conform to their chosen markets' regulations. In developing novel statistical methods to support diagnostic development, such as the use of adaptive design for sample size reassessment, it is paramount that the regulations, and associated guidance, do not preclude the proposed novel methodology. This review of legislation, official policy guidance, and standards across the UK, EU, and US aimed to identify regulatory requirements or restrictions relating to statistical methodology for diagnostics development.

Methods Data sources identified for legislation, official policy guidance, and standards included: *EUR-Lex*, *WestLaw UK*, *US Food and Drug Administration (FDA)*, *Lexis+*, *Policy Commons*, *Medical Device Co-ordination Group (MDCG)*, and the *British Standards Online Library*. These data sources were searched for records relating to medical diagnostic development. Search terms included: *Medical Device*, *In Vitro Diagnostic*, *Medical Diagnostic*, *Diagnostic*, and *IVD*. Identified records were double screened for inclusion, including a within document search for 25 key terms related to statistical requirements and diagnostic development. Identified terms were coded and relevant statistical requirements both mandatory and recommended, extracted.

Results This systematic review identified 2479 potential records, 540 of which met the inclusion criteria for data extraction, of which 139 had statistical requirements or recommendations related to medical diagnostic development. Mandatory requirements for specific

tests or conditions were identified across the three regions (Total: n =187, UK = 12, EU = 82, US = 93). Examples of requirements include minimum sample sizes and specific populations when demonstrating diagnostic accuracy in certain high-risk conditions. For example, the EU Common Technical Specifications require first line assays for anti-HIV1/2 to include 400 positive HIV-1 and 100 positive HIV-2 specimens, of which 40 are non-B subtypes, and 25 are 'same day' fresh serum. Whilst not mandatory, this review also identified recommendations for best practice in diagnostic development and trial design covering: *evidence requirements, statistical validity, study design, and study management.*

Conclusion Whilst mandatory statistical requirements exist for high-risk areas, thereby limiting the potential benefit of an adaptive trial due to mandating sample sizes, there remains a great opportunity for the development of novel methodologies and adaptive trial designs in medical diagnostics. This review will allow future development of a framework for designing adaptive trial in medical diagnostics, empowering statisticians and developers to improve efficiency whilst meeting regulatory requirements.

19: Calf Muscle Development in NICU Graduates Compared with Typically Developing Babies: An Analysis of Growth Trajectories using Linear Mixed Models

Alana Cavadino¹, Sian Williams^{2,3}, Malcolm Battin⁴, Ali Mirjalili⁵, Louise Pearce⁶, Amy Mulqueeney⁴, N. Susan Stott⁷

¹Epidemiology & Biostatistics, Faculty of Medical and Health Sciences, University of Auckland, New Zealand

²Curtin School of Allied Health, Faculty of Health Sciences, Curtin University, Australia

³Liggins Institute, University of Auckland, New Zealand

⁴Newborn Services, Starship Child Health, Auckland District Health Board, New Zealand

⁵Department of Anatomy and Medical Imaging, Faculty of Medical and Health Sciences, University of Auckland, New Zealand

⁶Auckland Children's Physiotherapy, Auckland, New Zealand

⁷Department of Surgery, Faculty of Medical and Health Sciences, University of Auckland, New Zealand

Background / Introduction

Preterm birth and Neonatal Intensive Care Unit (NICU) admission are related to adverse health consequences in early childhood and beyond. This study evaluated lower leg muscle growth and motor development in the first 12 months of life in NICU graduates compared

to typically developing (TD) infants.

Methods

A prospective, longitudinal study of infants born in Auckland, New Zealand, without complications and recruited from the community (TD), or discharged from a NICU (classed as intermediate-risk (NICU-IR) or higher-risk (NICU-HR) based on additional risk factors for adverse neurodevelopmental outcomes. Muscle volume and gross motor development were assessed at term-corrected ages 3-, 6- and 12-months (± 1 month). Linear mixed models with REML and Kenward-Roger small-sample adjustment were used to estimate trajectories in Triceps Surae muscle volume measurements (Medial Gastrocnemius, Lateral Gastrocnemius, Soleus, and total Triceps Surae). Models included random intercepts for individuals and slopes for term-corrected-age, and fixed effects for term-corrected-age (months), body side (left/right leg), group (TD/NICU-IR/NICU-HR), and sex. Non-linear terms and interactions (by-group and by-side) for term-corrected-age, and different variance-covariance structures were evaluated. Estimated group trajectories and marginal means at 3-, 6- and 12-months term-corrected-age were presented.

Results

Sixty-one infants were recruited; n=24 TD, n=14 NICU-IR, and n=23 NICU-HR. NICU infants had lower birthweight (1.7 ± 0.9 kg) and length (40.3 ± 6.2 cm) compared to TD infants (3.3 ± 0.5 kg; 51.1 ± 2.8 cm). COVID-19 restrictions meant some 6- and 12-month assessments occurred late, with variable timings. For muscle volume measures, there were significant term-corrected-age*group and (term-corrected-age) 2 *group interactions, indicating muscle growth trajectories over time differed by group (Medial Gastrocnemius, Lateral Gastrocnemius, Triceps Surae, $p < 0.001$; Soleus, $p = 0.04$). Negative correlations between random intercepts and slopes indicated lower muscle volume at 3-months term-corrected-age was associated with faster growth. Between 3-12 months term-corrected-age, Triceps Surae increased on average by 18.1cm^3 (95%CI: $16.1\text{-}20.2\text{cm}^3$), 13.3cm^3 ($10.6\text{-}16.0\text{cm}^3$) and 12.5cm^3 ($10.5\text{-}14.6\text{cm}^3$) in TD, NICU-IR, and NICU-HR infants, respectively. Soleus was smaller at 6- and 12-months term-corrected-age for both NICU groups, and Lateral Gastrocnemius was smaller at 12-months, term-corrected-age, for NICU-HR ($p < 0.001$). At 12-months term-corrected-age, raw Gross Motor Quotient scores were lower for NICU-HR ($p = 0.005$), and <10% of NICU infants were walking compared to 30% of TD.

Conclusion

Failure of typical Soleus growth over the first year contributed to a smaller Triceps Surae at 12-months term-corrected-age in NICU graduates. These findings add to the increasing body of evidence for an adverse impact of preterm birth and NICU stays on infant skeletal muscle growth.

20: Automating Report Generation with Stata: A Case Study of NORUSE

Maria Elstad

Helse Stavanger, Norway

Abstract

The Norwegian Service User Registry (NORUSE) is a comprehensive health registry utilized by Norwegian municipalities to document service recipients with substance abuse and/or mental health issues. The primary goal of NORUSE is to gather knowledge about the extent of services and the expected demand for services for this patient group. This data supports the formulation of municipal substance abuse policies, better decision-making regarding prioritization of user groups, and improved evaluation of service offerings. Nationally, the statistics contribute to the data foundation for shaping national policies for mental health and substance abuse work.

In 2024, we generated 64 automated municipality reports using VBA code in Excel. However, we have begun exploring the use of the Stata command `putdocx` for creating these reports. We are already using this for subgroup analysis, regional and national reports. This exploration highlights the potential of `putdocx` to streamline the process of generating detailed and consistent reports. Although we have also considered other software like Power BI, we found it less flexible compared to Stata, despite its superior graphing capabilities.

By employing `putdocx`, we can automate the creation of reports, which is particularly beneficial for municipalities that receive community-specific reports shortly after data collection. Additionally, Helse Stavanger produces regional and national reports, further leveraging the efficiency of automated report generation. The integration of `putdocx` in our reporting workflow enhances the accuracy and timeliness of data presentation, supporting better decision-making and policy formulation.

As we consider employing this method more broadly, we anticipate significant improvements in our ability to provide clear snapshots of users' situations based on the latest contact status. This tool contributes significantly to the ongoing efforts to improve service delivery for individuals with substance abuse and mental health challenges. The flexibility and scalability of `putdocx` make it a promising solution for our future reporting needs.

21: Maternal Mortality Rate in Sudan 2020: Causes of Death, Obstetric Characteristics and Territorial Disparity, Using Statistical Analysis.

Mohammed Abdu Mudawi

freelancer (1Senior Statistician, Health Information System and Biostatistics Specialist) (Health (Health Information System))

Abstract

Maternal mortality in general is deaths associated with pregnancy. Maternal mortality is one of crucial of social determinants of health and sociodemographic to measure and evaluate the quality of health care services (Antenatal Care Services), and reflects to the strength of health system in general, although Sudan was among the first countries in the Arab and Africa region, which conducted the (demographical health survey ¹⁹⁸⁹, safe motherhood survey ¹⁹⁹⁰, Sudan household health survey ^{2006 & 2010} and multiple indicators cluster survey ²⁰¹⁴), but the last survey that conducted and included the maternal mortality rate was been in 2010 and the maternal mortality rate was been 216 per 100000 live birth, due the instability situation (Sudanese Revolution 2018 and political situation), the sixth multiple indicators cluster survey (MICS 6th) 2018 was not conducted.

The paper focus and illustrates the estimation the Maternal Mortality Rate - MMR in Sudan by causes of diseases, place of deaths, obstetric characteristics and territorial disparity, the data were collected from the Federal Ministry of Health (Annual Statistical Report and Maternal Mortality Deaths Surveillance) for the year of 2020.

The Maternal Mortality Rate – MMR in the country was (278.7 per 100000 live birth) in 2020, and the higher Maternal Mortality Rate in the East Darfur state was (1531.8 per 100000 live birth), and most maternal deaths was happened due the Obstetric Hemorrhage by (35%), and (45%) of maternal deaths in age between (20 - 30), 508 deaths (62%) happened out of antenatal care and ANC follow up services. West Kordofan state was most state registered the maternal deaths by (10% of deaths of all states), and the most maternal deaths was happened in health facilities by (82% of deaths according of place) it was more than deaths at home and in road. The Maternal Mortality Rate higher in 2020 than the last in SHHS survey 2010 was been 216 per 100000 live birth.

22: Community-Based Health Screening Attendance and All-Cause Mortality in Rural South Africa: A Causal Analysis

Faith Magut¹, Stephen Olivier¹, Ariane Sesseg¹, Lusanda Mazibuko¹, Jacob Busang¹, Dickman Gareta^{1,6}, Kobus Herbst^{1,5}, Kathy Baisely^{3,1}, Mark Siedner^{1,2,4}

¹Africa Health Research Institute (AHRI), South Africa

²Massachusetts General Hospital, Boston, Massachusetts, United States of America

³London School of Hygiene & Tropical Medicine, Keppel Street, London, UK

⁴University of KwaZulu-Natal, Durban, South Africa

⁵DSI-SAMRC South African Population Research Infrastructure Population Infrastructure Network, Durban, South Africa

⁶Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland

Background South Africa is moving from a period marked by high mortality from HIV and tuberculosis(TB) to one characterised by a growing burden of non-communicable diseases. Community health fairs help to diagnose and refer individuals with chronic diseases in underserved areas. However, their impact on morbidity and all-cause mortality is unknown.

Methods We enrolled individuals 15 years and older in the Africa Health Research Institute Health Demographic Health Surveillance area in rural KwaZulu-Natal to a community-based health fair screening and referral program (Vukuzazi). Testing was performed for HIV, TB, hypertension and diabetes. Those with positive results were visited at home for results provision and referral to local clinics.

All Individuals in the area were followed longitudinally through routine household surveillance to detect deaths. We used directed acyclic graphs to identify the following confounders of the association between health-fair attendance and mortality: age, sex, educational attainment, employment, household socio-economic status and prior healthcare-seeking behavior.

To estimate the effect of Vukuzazi health fair attendance on all-cause mortality, we first estimated inverse probability of treatment weights (IPTW) for health fair attendance, then applied weighted Kaplan-Meier analysis to compare survival and weighted Cox regression to estimate hazard ratios and marginal risk differences. We conducted a sensitivity analysis where we excluded deaths due to external factors (e.g. injuries) that would not be expected to be prevented by health fair attendance

Results A total of 18,041 individuals (50.0% of those eligible) attended Vukuzazi health fairs. Compared to non-attenders, attenders were more likely to be women (68% vs 49%), older (median 37 vs 31 years), unemployed (37% vs 20%) and more likely to have accessed health care in the past year (53% vs 33%). Individuals were observed after health fairs for a median

of 4.0 years (IQR 3.7 - 4.2 years) comprising a total of 127,625 person years. The crude mortality rate was 12.14(11.54-12.76) per 1000 person years. In weighted Kaplan-Meier analysis, attenders had better survival compared to non-attenders. In the IPTW adjusted models, Vukuzazi health fair attendance was associated with a 25% reduction in the hazard of all-cause mortality ($HR=0.75$, 95%CI: 0.67, 0.84), corresponding to a 1.5% absolute reduction in mortality over five years. Findings were similar in sensitivity analysis.

Discussion Participation in a community-based health fair was associated with a reduction in 5-year all-cause mortality. The integration of health fairs with referral practices into standard healthcare delivery within rural areas may be an effective strategy to improve health outcomes.

23: Reducing Uncertainty in Fertility Meta-Analysis: A Multivariate Approach to Clinical Pregnancy and Live Birth Outcomes

Mahru Ahmad, Jack Wilkinson, Andy Vail

University of Manchester, United Kingdom

Background Meta-analyses of assisted reproductive technology (ART) trials commonly assess clinical pregnancy and live birth as separate outcomes, despite their hierarchical dependency. Many trials report pregnancy but not live birth, limiting the applicability of univariate meta-analyses for live birth outcomes. This can lead to imprecise estimates and uncertainty about intervention effectiveness. Multivariate meta-analysis (MVMA) offers a potential solution by jointly modelling related outcomes, maximizing the use of available data and improving statistical precision.

Objectives

This study aims to investigate whether Multivariate Meta Analysis (MVMA) provides a more reliable estimation of live birth outcomes compared to traditional univariate meta-analysis. Specifically, we:

1. **Construct an MVMA model** incorporating both clinical pregnancy and live birth outcomes using data from systematic reviews of ART trials (2020–2021).
2. **Compare MVMA with univariate approaches**, evaluating the extent to which MVMA improves precision and whether this would lead to different inferences. .
3. **Explore different correlation structures** between clinical pregnancy and live birth,

assessing their impact on effect estimates.

Methods

Systematic review data from the Cochrane systematic reviews (2020–2021) will be extracted, including trial-level counts of clinical pregnancies and live births for treatment and control groups. MVMA models will be implemented using various correlation assumptions, as well as the use of the Wei and Higgins method to account for the relationship between outcomes. The study will assess the performance of MVMA versus univariate meta-analysis by comparing uncertainty in effect estimates and methodological implications.

Results

This study will provide insights into whether MVMA can enhance the precision of live birth effect estimates, making better use of incomplete ART trial data. By improving the analysis of imperfectly reported data, this study aims to reduce the considerable uncertainty surrounding many fertility interventions. The results will provide insights into whether MVMA leads to more precise effect estimates compared to univariate methods. The findings will be available at the time of the presentation and will help determine the extent to which MVMA can enhance statistical power when live birth data is incomplete. This work will contribute to methodological advancements in fertility research by optimising the use of available trial data and improving the reliability of conclusions drawn from ART studies.

24: Causal Discovery for Multi-Cohort Studies

Christine Bang¹, Vanessa Didelez^{2,3}

¹University of Copenhagen

²Leibniz Institute for Prevention Research and Epidemiology - BIPS

³University of Bremen

Causal discovery methods aim to learn causal structures in a data-driven way. The availability of multiple overlapping cohort datasets enables us to learn causal pathways over an entire lifespan. Evidence of such pathways may be highly valuable, e.g. in life course epidemiology. No previous causal discovery methods tailored to this framework exist. We show how to adapt an existing causal discovery algorithm for overlapping datasets to account for the time structure embedded in cohort data. In particular, we show that this strengthens the method in multiple aspects.

We consider causal discovery methods that recover causal structures from (conditional) independencies in a given set of variables. Multiple causal structures may induce the same dependence structure and form an equivalence class. Without additional, stronger assump-

tions, it is usually not possible to recover more than the equivalence class; i.e. we cannot identify all causal directions. Moreover, when combining multiple datasets, if some variables are never measured jointly their (conditional in-)dependence is by construction unknown. Then, we cannot even identify the equivalence class. Hence, constraint-based causal discovery for multiple datasets suffers from two types of obstacles for identification.

Time structured data induces a partial causal ordering of the variables, which we refer to as tiered background knowledge. It is easy to see that tiered background knowledge improves the identifiability of causal directions. Additionally, we show that tiered background knowledge also improves the (partial) identifiability of the equivalence class, which is not trivial. We provide theoretical results on the informativeness as well as theoretical guarantees of the algorithm. Finally, we provide detailed examples that illustrate how the algorithm proceeds, as well as examples of cases where tiered background knowledge increases the level of informativeness.

25: Extension of Causal Interaction Estimation Techniques Through Integration of Machine Learning Algorithms

A F M Tahsin Shahriar, AHM Mahbub-ul Latif

University of Dhaka, Bangladesh, People's Republic of

This study explores the challenges of causal interaction analysis, particularly in public health and policy evaluation, where understanding how multiple exposures influence outcomes is crucial. Identifying these interactions is complex due to unobserved confounding, measurement errors, and high-dimensional datasets. Traditional econometric methods, while widely used, often rely on strong assumptions that may not hold in complex real-world scenarios.

This study reviews established causal inference methods, including Difference-in-Differences (DiD), Changes-in-Changes (CiC), and matching. These methods have limitations, particularly in handling high-dimensional data and complex interactions. To address these challenges, this research investigates an alternative approach using machine learning models, specifically Causal Forests and Bayesian Additive Regression Trees (BART), to estimate causal interactions. These models are used to obtain Conditional Average Treatment Effect (CATE) estimates, which are then used to compute the Average Treatment Effect on the Treated (ATET). However, these methods did not consistently outperform traditional methods in simulations, especially with smaller samples.

A key contribution of this study is the development of causal mixture methods, which in-

tegrate the adaptability of machine learning algorithms, like Gradient Boosting Machines (GBM) and Random Forests (RF), for first-stage estimation with the interpretability and robustness of traditional econometric frameworks, such as Difference-in-Differences (DiD), to enhance resilience to unmeasured confounding and measurement errors. This approach involves first estimating propensity scores using machine learning methods to capture complex relationships between covariates and treatment assignment. These estimated propensity scores are then integrated into the standard DiD model to improve covariate balance and comparability between treated and control groups, mitigating selection bias and enhancing the robustness of causal estimates. This approach aligns with modern econometric frameworks like Double Machine Learning (DML).

Simulation studies were conducted to assess the performance of various causal inference methods. Data were generated with varying levels of noise to examine the impact of measurement error. The mixture methods, integrating ML-based propensity scores with DiD regression, produced unbiased estimates, demonstrating robustness to measurement error.

In summary, this study advances the field of causal inference by: (i) presenting a detailed comparative analysis of econometric and machine learning-based methods, (ii) proposing causal mixture models that integrate machine learning for robust first-stage estimation, and (iii) comparing bias through simulations. These contributions provide researchers with practical tools and a stronger theoretical foundation for addressing challenges in causal interaction analysis, particularly in high-dimensional and complex settings, ensuring more reliable and interpretable conclusions for decision-making in public health and policy research.

26: Embrace Variety, Find Balance: Integrating Clinical Trial and External Data Using Causal Inference Methods

Rima Izem¹, Yuan Tian², Robin Dunn³, Weihua Cao³

¹Novartis Pharma AG, Switzerland

²China Novartis Institutes for BioMedical Research Co., Ltd.

³Novartis Pharmaceuticals Corporation, USA

Integrating information from multiple sources is important for multiple stakeholders in the development of pharmaceutical products. For example, augmenting the control arm of a randomized controlled trial with external data from previously conducted trials can inform internal decision-making in early development or expedite development in small populations with unmet medical need. Also, leveraging external controls from a disease registry to a

single arm trial can make it possible to estimate the comparative treatment effect of the study drug when a randomized comparison is unfeasible or unethical. The main challenge in this data integration is assessing potential biases, due to between-source differences, and minimizing or mitigating these biases in the integrated design and analysis.

This presentation proposes the use of a workflow implementing propensity score methods, developed in observational data, when estimating treatment effects from multiple data sources with individual-level data. First, causal inference thinking can help identify the causal estimand, establish the underlying assumptions, and focus the assessment of between-source heterogeneity on key variables. The use of target trial emulation and balance diagnostics can identify the relevant subset in the external data, assess the extent of adjustment needed, evaluate the plausibility of important assumptions, such as positivity, and assess adequacy of propensity score adjustment. Lastly, for fit-for-purpose external data, a variety of methods can leverage the propensity score to estimate the treatment effect. Our presentation will share practical considerations at each step of the workflow and illustrate its use with case studies and simulated data from pharmaceutical development.

27: Revisiting Subgroup Analysis: A Reflection on Health Disparities using Conditional Independence

Nia Kang, Tibor Schuster

McGill University, Canada

Introduction Comparative assessment is deeply ingrained in human nature to answer cause and effect questions. It is also an important feature of methodological rigour, underlying many research designs including randomized controlled trials, epidemiological studies and population-level evaluations for informing health policy. Programs that aim at addressing health disparities often rely on comparisons of health indicators across predefined subpopulations (i.e., groups distinguished by fixed socio-demographic characteristics), rather than by theoretically assignable exposures or interventions.

Although tailoring health policy implications to such subgroups may seem reasonable, this approach risks oversimplification, as the intersectional nature of socio-demographic factors can obscure those with the greatest need, rendering population-level interventions derived from such analyses less effective.

Methods Using principles from probability theory, we define health parity as the stochastic independence between one or more health indicators and any subdivision of the population

conditional on confounding factors. We consider the presence of two or more group-defining features that may intersect within and across subpopulations. We further assume the availability of a program or policy P that has a positive causal impact on the health indicator(s) under study but has limited resource allocation.

Using Bayes' theorem, we derived a target function that factorizes the tradeoff between decreasing subgroup-specific health disparities and lowering the marginal prevalence of a poor health outcome given practical constraints such as resource availability. We conducted extensive Monte Carlo simulation studies to demonstrate how the proposed function can help identify the most optimal P in terms of maximizing health parity. Factors considered in the simulations are the degree of impact of P , resource availability, number and prevalence of population subgroups, and varying distributions of health outcomes.

Results/Conclusion: The proposed functional approach demonstrated utility in assessing the effectiveness of health programs and policies aimed at maximizing health parity. Although subpopulations defined based on sociodemographic features provide an easy ground for conventional comparative assessment, they may have limited capacity to inform the most effective health policies. Indeed, our findings imply that comparative subgroup analysis should be supplemented with marginal outcome distributions by leveraging the proposed target function approach.

28: Comparison of Multiple Imputation Approaches for Skewed Outcomes in Randomised Trials: a Simulation Study

Jingya Zhao, Gareth Ambler, Baptiste Leurent

University College London, United Kingdom

Introduction Missing outcome data is a common issue in trials, leading to information loss and potential bias. Multiple imputation (MI) is commonly used to impute missing data; one advantage is that it can include additional predictors of 'missingness' that are not in the analysis model. However, standard MI methods assume normality for continuous variables, which is often violated in practice, e.g. healthcare costs are typically highly skewed. Alternative MI approaches, involving Predictive Mean Matching (PMM) or log transformations, have been proposed for handling skewed variables. Using simulation, we compare different methods for imputing missing values of skewed outcome variables in randomised trials.

Methods We simulated trial data with two treatment arms and correlated skewed baseline and follow-up variables. We considered three different missing data mechanisms for

the follow-up variable: missing completely at random (MCAR), missingness associated with treatment arm (MAR-T), and missingness associated with baseline (MAR-B). We compared seven methods: Complete Case Analysis (CCA), Multivariate Normal Imputation (MVN), Multiple Imputation by Chained Equations (MICE), and Predictive Mean Matching (PMM), along with log-transformed versions (LogMVN, LogMICE, and LogPMM) which perform imputation on the log-transformed variables. Assessment of performance focused on bias and confidence interval (CI) coverage when estimating the mean difference between arms. These methods were also applied to the analysis of a healthcare costs trial dataset.

Results The simulation results showed that LogMVN and LogMICE typically outperformed other methods. MVN and MICE also performed well under MCAR and MAR-T but had poor performance under MAR-B. PMM and LogPMM generally performed poorly, often showing under-coverage. CCA performed well under MCAR but not under MAR mechanisms. When applied to the trial dataset, PMM and LogPMM produced point estimates similar to that of CCA, with the narrowest CIs. Conversely, LogMVN and LogMICE yield higher point estimates, along with the widest CIs. Additional simulations are being performed to explore further results under different outcome distributions, missing data mechanism and sample sizes.

Conclusion Our results suggest that a log transformation before MI strategy might be useful for handling skewed variables (although non-positive values need careful handling). The performance of MVN and MICE depends on the specific missingness mechanism, and the PMM method cannot be recommended. However, further evaluation alternative data generation mechanisms are needed.

29: Assessing the Effect of Drug Adherence on Longitudinal Clinical Outcomes: A Comparison of Instrumental Variable and Inverse Probability Weighting Methods.

Xiaoran Liang¹, Deniz Türkmen¹, Jane A H Masoli^{1,2}, Luke C Pilling¹, Jack Bowden^{1,3}

¹University of Exeter, United Kingdom

²RoyalDevon University Healthcare NHS Foundation Trust, Exeter, United Kingdom

³Novo Nordisk Research Centre (NNRCC), Oxford, United Kingdom

Background Drug adherence refers to the degree to which patients comply with prescribed therapeutic regimens when taking medications and high adherence is essential for ensuring the expected efficacy of pharmacological treatments. However, in routine care settings, low adherence is a major obstacle that frustrates this desired process. For instance, real-world studies report that adherence to commonly provided statin therapy can drop below 50%

within the first year of treatment, which is substantially lower than observed in the controlled trials that led to their original approval. **Method:** In this paper we discuss the use of longitudinal causal modelling to estimate the time-varying causal effects of adherence on patients' health outcomes over a sustained period. The goal of such analyses is to provide a means for quantifying the impact of interventions to improve adherence on long-term health. If a meaningfully large difference is estimated by such an analysis, the natural focus can then shift to deciding how to realize such an intervention in a cost-effective manner. Two estimation approaches, Inverse Probability Weighting (IPW) and Instrumental Variables (IV), have been proposed in the 'Estimand framework' literature to adjust for non-adherence in randomized clinical trials, where non-adherence is viewed as an intercurrent event. We refine and adapt these methods to assess long-term adherence in the observational data setting, which differs in several key respects compared to a clinical trial: Firstly, an absence of overt randomization to treatment, secondly, adherence and longitudinal outcomes only being available in those who are treated. We clarify the assumptions each method makes and assess the statistical properties of each approach using Monte Carlo simulation as well real data examples on statin use for LDL cholesterol control and metformin use for HbA1c control taken from primary care data in UK Biobank.

Results The findings from our simulations align with theoretical expectations. The IV method effectively accounts for time-varying observed and unobserved confounders but relies on strong, valid instruments and additional parametric assumptions on the causal effects. In contrast, the IPW method addresses observed confounders without requiring additional assumptions but remains susceptib

30: Compliance Between Different Anthropometric Indexes Reflecting Nutritional Status in Women with Polycystic Ovary Syndrome

Aleksander J. Owczarek¹, Marta Kochanowicz², Paweł Madej², Magdalena Olszanecka-Glinianowicz¹

¹Health Promotion and Obesity Management Unit, Department of Pathophysiology, Faculty of Medical Sciences in Katowice, Medical University of Silesia in Katowice, Poland

²Department of Gynecological Endocrinology, Faculty of Medical Sciences in Katowice, Medical University of Silesia in Katowice, Poland

Background Obesity (mainly diagnosed based on the body mass index – BMI) is the main risk factor for developing polycystic ovary syndrome (PCOS). Based on BMI not all women with PCOS are diagnosed with obesity. However, BMI does not assess visceral fat deposits that play a key role in the pathogenesis of PCOS. Thus, there is a constant search for

anthropometric indicators that allow the assessment of visceral fat deposits. This study aimed to assess the comparison of various anthropometric indicators for the diagnosis of excessive fat deposits.

Methods Based on body mass, height, waist, and hip circumference eleven indexes were calculated: BMI, waist-to-hip ratio (WHR), waist-to-height ratio (WHtR), waist-to-hip-to-height ratio (WHHR), body adiposity index (BAI), a body shape index (ABSI), body roundness index(BRI), weight-adjusted waist index (WWI), abdominal volume index (AVI), CI (Conicity index), Roher or corpulence index (RI). To compare indexes with each other based on the Passing-Bablok (PB) regression, they were scaled to range [0,1]. The serum lipid profile total cholesterol, LDL and HDL cholesterol, triglycerides) as well as triglyceride-glucose index (TyG) were also determined.

Results The study group encloses 611 women with diagnosed PCOS, with a mean age of 26.3 ± 4.8 (range: 17 – 43) years. There were positive significant linear correlations between indexes (ranging from 0.08 to 0.99), except between ABSI versus BAI and RI. Overall, 55 comparisons between indexes were done with PB regression regarding intercept and slope. Apart from the comparison between BMI vs RI, WHHR vs WWI, and WHR vs ABSI, all methods were different from each other regarding the intercept (ranging from -0.28 to 0.24). Taking slope into consideration, 21 (38.2%) comparisons yielded slopes that did not differ significantly from 1. The highest positive slope value of 1.42 (95% CI: 1.27 – 1.57) was noted in the comparison between WWI vs BAI. The lowest value of 0.72 was noted in the comparison between BAI vs WHtR (95% CI: 0.68 – 0.77). Indexes the most consistent with each other were WHR vs WHtR and ABSI, BRI vs RI, and BMI vs BRI and RI. The highest significant correlations with lipid profile were observed for WHtR while the lowest for ABSI.

Conclusions Individual anthropometric indexes are not equivalent to each other. Assessment of the level of nutrition using different indicators may lead to over- or underdiagnosis of obesity among women with PCOS.

31: Effectiveness of Different Macronutrient Composition Diets on Weight Loss and Blood Pressure. a Network Meta-Analysis

Katerina Nikitara¹, Anna-Bettina Haidich², Meropi Kontogianni³, Vasiliki Bountziouka¹

¹Computer Simulation, Genomics and Data Analysis Laboratory, Department of Food Science and Nutrition, School of the Environment, University of the Aegean

²Department of Hygiene, Social-Preventive Medicine and Medical Statistics, School of Medicine, Faculty of Health Sciences, Aristotle University of Thessaloniki

³Department of Nutrition and Dietetics, School of Health Sciences and Education, Harokopio University

Background The scientific evidence surrounding the effectiveness of macronutrient composition on weight loss and reduction of Systolic Blood Pressure (SBP) and Diastolic Blood Pressure (DBP) is conflicting. Advanced analytical methods can be used to examine the effects of different macronutrient compositions. This study explored the effectiveness of diets with different macronutrient compositions for weight loss and blood pressure through network meta-analysis (NWMA).

Methods A systematic review was conducted by retrieving studies from five bibliographic databases (January 2013 to May 31, 2023). The study population included adults at high risk for cardiovascular diseases, while the outcomes assessed involved markers of glycemic control, obesity, dyslipidemia, and inflammation. Specifically, in the present study, the outcomes of interest were the mean difference in Body Mass Index (BMI), Waist Circumference (WC), SBP and DBP before and after the intervention. The reference diet used for BMI and WC was low-fat (<30%), moderate-carbohydrate (45-60%), and high-protein (19-40%) (LFMCHP) and for the SBP and DBP, low-fat, moderate-carbohydrate, and moderate-protein (10-18%) (LFMCMP), according to the reference diets used in the studies included for each outcome.

Results Ten studies ($n=1,008$ individuals) were included in NWMA for BMI, six ($n=835$) for WC, and seven ($n=1,103$) for SBP and DBP. The random effect model was used in the NWMA. Results revealed that, compared to the reference diet (LFMCHP), only the high-fat (36-60%), low-carbohydrate (26-44%), high-protein (HFLCHP) diet demonstrated a greater reduction in BMI after the intervention by 0.32 kg/m^2 (95% CI: -0.34; -0.30, $I^2=0\%$, $p\text{-value}<0.001$). Additionally, the highest ranking in terms of certainty of effectiveness was observed for the high-fat, very low-carbohydrate (<26%), very high-protein (>40%) diet (HFVLCVHP) ($P\text{-score}: 0.71$) compared to other interventions, followed by the HFLCHP diet ($P\text{-score}: 0.63$) and the high-fat, low-carbohydrate, moderate-protein diet (HFLCMP) ($P\text{-score}: 0.59$). Non-significant results were found for WC, SBP, and DBP.

Conclusion This NWMA suggests that high-fat, low-carbohydrate, high-protein diets may be more effective for BMI reduction, while no significant effects were observed for blood pressure. These findings highlight the potential role of macronutrient composition in weight management but indicate the need for further research to clarify its impact on other cardiometabolic outcomes.

32: Going from Methodological Research to Methods Guidance: The STAndards for the Development REseArch Methods Guidance (STREAM) Initiative

Malena Chiaborelli, Julian Hirt, Matthias Briel, Stefan Schodelmaier

University Hospital Basel, Switzerland

Background Health researchers need clear and trustworthy methods guidance (e.g. tutorials on handling baseline missing data in trials; best practice regarding calibration of prediction models) to help them plan, conduct, and analyse their studies. Methodological research (based on logic, simulation, or empirical studies) can sensibly inform methods guidance. How to go from methodological research to methods guidance, however, is currently unclear. A new initiative (Standards for the Development of Research Methods Guidance, STREAM) aims to develop a structured process to connect methodological research with methods guidance.

Methods STREAM includes a series of studies: 1) a scoping review of existing standards to develop methods guidance, 2) a meta-study to assess the current practice of methods guidance development, 3) an interview study to understand the needs of health researchers who use methods guidance, 4) a consensus study to develop standards for methods guidance development, and 5) user testing of these standards in ongoing guidance development projects.

Results At the conference, we will present the overall initiative and results of the first two studies. The scoping review identified 6 articles addressing the development of methods guidance. Of those, 1 mentioned methodological research (specifically: empirical studies) as an input for guidance development, without specifying a process. None of the included articles mentioned simulation studies as an input. For the meta-study, we reviewed 1202 methods guidance articles, most published after 2018. Of those, 347 reported a development process: 156 (45%) performed a systematic review of the methodological literature, 93 (27%) a consensus process, 71 (20%) user-testing, 43 (12%) empirical studies, and 36 (10%) simulation studies.

Impact: The two initial studies of the STREAM initiative reveal that the literature addressing the development of methods guidance is scarce and limited and that methods guidance articles rarely report a development process. Guidance developers use varying ad hoc approaches to create guidance and rarely seek input from their users (health researchers). The findings suggest that current methods guidance could be improved to make it more helpful for health researchers and better support the production of high-quality evidence. The new standards for the development of research methods guidance will provide explicit solutions to these challenges.

33: Effectiveness of a Skill Check Sheet for Registered Dietitians: A Cluster Randomized Controlled Trial Protocol

Misa Adachi^{1,2}, Asuka Suzuki², Kazue Yamaoka^{1,3}, Mariko Watanabe⁴, Toshiro Tango^{1,5}

¹Nutrition Support Network LLC, Sagamihara, Japan

²Teikyo University Graduate School of Public Health, Japan

³Tetsuyu Clinical Research Center, Tetsuyu Institute Medical Corporation, Tokyo, Japan

⁴Showa Women's University, Tokyo, Japan

⁵Center for Medical Statistics, Tokyo, Japan

Introduction:

Registered dietitians (RDs) play a critical role in promoting lifestyle improvements through evidence-based nutrition interventions. To enhance RD competencies in nutrition education, we developed a Skill Check Sheet (SCS) designed to support self-assessment and skill improvement. A preliminary single-group intervention study (3 months) suggested that SCS might effectively improve RD skills. This study aims to evaluate its effectiveness in reducing glycated hemoglobin (HbA1c) levels among patients with type 2 diabetes (T2D) by conducting a cluster randomized controlled trial (cRCT). The intervention compares a validated nutrition education program, the SILE program (Adachi et al., 2017), with an enhanced version incorporating the SCS (SILE+SCS).

Methods and Results This 4-month cRCT will randomly assign RDs to one of two intervention arms (SILE+SCS vs. SILE). Each RD will manage seven T2D patients aged 20–80 years. The primary outcome is the change in HbA1c from baseline. The intervention effect will be assessed using an intention-to-treat (ITT) analysis with a generalized linear mixed-effects model, adjusting for covariates.

The sample size calculation was based on previous studies and preliminary data, assuming a standardized mean difference (SMD) of 0.33, an intraclass correlation coefficient (ICC) of 0.01, a two-sided significance level of 5%, and 80% power, with seven patients per RD cluster. This resulted in a required sample of 21 RDs per group. Accounting for a 10% dropout rate, the final target is 23 RDs per group, totaling 322 patients.

Conclusion Preliminary findings suggest that SCS may enhance RD skills in nutrition education. This cRCT will rigorously evaluate its effectiveness, ultimately aiming to contribute to the prevention and management of lifestyle-related diseases.

Reference

Adachi M, Yamaoka K, Watanabe M, et al. Does the behavioural type-specific approach for type 2 diabetes promote changes in lifestyle? Protocol of a cluster randomised trial in Japan. BMJ Open 2017;7:e017838. doi:10.1136/bmjopen-2017-017838 Adachi et al., 2017.

34: Bootstrap-Based Approaches for Inference on the Total Deviation Index in Agreement Studies with Replicates

Anna Felip-Badia¹, Josep L Carrasco², Sara Perez-Jaume^{1,2}

¹BiMaU, Sant Joan de Déu Pediatric Cancer Center Barcelona, Spain

²Department of Basic Clinical Practice, Universitat de Barcelona, Spain

Introduction The total deviation index (TDI) is an unscaled statistical measure used to evaluate the deviation between paired quantitative measurements when assessing the extent of agreement between different raters. It describes a boundary such that a large specified proportion of the differences in paired measurements are within the boundary (Lin, 2000). The inference of the TDI involves the estimation of a $100(1-\alpha)\%$ upper bound (UB), where α is the significance level. Some methods to estimate the TDI and the UB have been proposed (Choudhary, 2008, 2010; Escaramis, 2010). In 2015, Perez-Jaume and Carrasco (P-J&C) proposed a non-parametric method that estimates the TDI as a quantile of the absolute value of the within-subject differences between raters and bootstraps them with two strategies to estimate the UB. Our goal is to assess an alternative bootstrap approach when estimating the UB using P-J&C's method, and to compare its performance as well as the one of the TDI estimates to that of the already existing methods in the literature.

Methods

We consider two non-parametric bootstrap approaches for studies with replicates: the bootstrap of the within-subject differences and an alternative approach of a cluster bootstrap at subject level. We also consider four strategies to estimate the UB: the ones based on the basic percentile and the normal distribution from P-J&C and two additional ones based on empirical quantiles and BC_a confidence limits. This leads to eight different ways of UB estimation. We implement all the above-mentioned methods to estimate the TDI and the bootstrap-based approaches for inference in an R package and conduct a simulation study to compare the performance of all the methodologies considered in this work. Furthermore, we apply them to a real case dataset.

Results

All the methods exhibit a tendency to overestimate the TDI except for Choudhary's 2010 method that seems to underestimate it in all combinations considered in the simulation study.

The bias and the mean squared error is reduced when the sample size is increased for all methods, indicating consistent asymptotic properties. Regarding the empirical coverages, the cluster bootstrap approach gives values closer to the expected 95% than the ones from the bootstrap of the within-subject differences. Finally, under real data with replicates all techniques provided similar estimates with the BC_a strategy resulting in slightly higher UBs in most cases.

Conclusion

In studies with replicates, when applying bootstrapping to estimate the UB using the P-J&C estimator, the cluster bootstrap approach is recommended.

35: Baseline Treatment Group Adjustment in the BEST Study, a Longitudinal Randomised Controlled Trial.

Robin Young¹, Alex McConnachie¹, Helen Minnis²

¹Robertson Centre for Biostatistics, University of Glasgow, United Kingdom

²Centre for Developmental Adversity and Resilience (CeDAR), University of Glasgow, United Kingdom

In an RCT with measurements of the outcome variable at baseline and one or more follow up visits, a linear mixed effects regression model can be used. Due to randomisation it would be expected that there is no difference between treatment groups at baseline, and so a model term for treatment effect at baseline can be omitted. It has been shown that such a “constrained baseline analysis” would have more power than if a term for the baseline treatment effect is included in analysis models¹.

The BEST² trial was an RCT assessing the impact of the New Orleans Intervention Model on children entering foster care in the UK, with measurement of outcomes at baseline and two follow up visits. As a result of practical and legal considerations relating to the setting of the study, over the 10 year duration of the trial there were three separate schedules of recruitment: (1) Consent first followed by baseline measures and then randomisation (2) Randomisation followed by consent then baseline (3) Consent followed by randomisation then baseline. As not all participants were recruited with randomisation occurring after baseline, it could not be guaranteed prior to unblinding at the end of the study that the treatment groups were balanced at baseline for the primary outcome. A pre-defined statistical analysis plan for the study therefore took the approach to include a term for the treatment effect at baseline to account for any unexpected differences.

At the conclusion of the trial, there was some degree of difference at baseline between the unblinded treatment groups for the primary outcome, and as a result the choice to include a term for this in the primary analysis model appeared justified. Using the data from the trial in combination with simulations, we will show that there are scenarios where due to study design, or to account for high variability in outcome measures, including the baseline treatment effect may be relevant to consider as either the primary model or as a sensitivity to constrained baseline analysis.

References [1] Coffman CJ, Edelman D, Woolson RF, To condition or not condition? Analysing 'change' in longitudinal randomised controlled trials. *BMJ Open* 2016;6:e013096. doi: 10.1136/bmjopen-2016-013096

[2] BEST [Accepted Nature medicine]

36: The Subtle Yet Impactful Choices in Procedure to Conduct Matching-Adjusted Indirect Comparison - Insights from Simulation

Gregory Chen¹, Micahel Seo², Isaac Gravestock²

¹MSD, Switzerland

²Roche, Switzerland

Population-adjusted indirect treatment comparisons (ITCs) play a crucial role in clinical biostatistics, particularly in the health technology assessment (HTA) space. Demonstrating the comparative effectiveness of an investigational treatment against standard-of-care comparators is essential for both clinical and economic decision-making in reimbursement submissions. However, head-to-head randomized trials for payer-interested comparators are often unavailable at the time of a HTA submission, necessitating the use of indirect comparison methods.

When only aggregate data (AgD) are available for a comparator, the Matching-Adjusted Indirect Comparison (MAIC) method, originally introduced by Signorovitch, has become the go-to approach. Over time, variations and refinements have been introduced in both research and practice. This study conducts a simulation-based evaluation of the bias and relative efficiency of different MAIC estimators for the average treatment effect among treated (ATT), along with an assessment of confidence interval (CI) coverage based on asymptotic derivations, robust variance estimators, and bootstrap methods.

The simulation utilizes {maicplus} R package and is designed to generate insights for both

binary and time to event endpoints. The primary focus is on unanchored ITCs, with a secondary analysis of anchored comparisons to assess the robustness of findings. The study examines performance across various scenarios, including different sample sizes, true event rates, and degrees of prognostic factor overlap. Additionally, we investigate the impact of including non-prognostic factors, omitting key confounders, and interactions between these factors. To further contextualize MAIC findings, we incorporate inverse probability of treatment weighting (IPTW) estimators, quantifying the trade-offs in performance metrics when individual patient data (IPD) for the comparator arm are unavailable.

The findings from this study will provide critical insights into the feasibility, reliability, and trade-offs of population-adjusted ITCs, offering guidance on best practices and methodological considerations in comparative effectiveness research.

37: Utility-Based Design: An Improved Approach to Jointly Analyze Efficacy and Safety in Randomized Comparative Trials

Patrick Djidel, Armand Chouzy, Pierre Colin

Bristol Myers Squibb, Switzerland

Introduction In randomized clinical trials, multiple endpoints are evaluated to assess new treatments, focusing on both efficacy and safety. Traditional oncology study designs often rely on a single primary endpoint, which can overlook other important objectives. Various frameworks, such as those proposed by Murray, Kavelaars, and Park, incorporate multivariate outcomes to improve decision-making by considering the risk-benefit tradeoff. We propose a utility-based design tool, extending Murray's approach, that accounts for the correlation between efficacy, safety and the cause of death (due to disease progression vs. fatal adverse event).

Methods The proposed statistical framework is based on a joint probit model as follows: the clinical endpoints are considered categorical (e.g. toxicity grade and objective response rate) and a composite endpoint is derived based on combinations of both safety and efficacy categories and numerical utilities. The utility matrix is obtained via a consensus among clinical trial physicians. Then, to evaluate the treatment effect, we calculate the mean joint probabilities via a joint probit model and combine them with the utility matrix. To support decision-making, a formal test is derived to analyze the improvement of the utility score due to the treatment effect.

Results We provide a statistical tool to efficiently compare treatment arms from randomized trials and evaluate the efficacy/safety trade-off. A statistical test and a target sample size calculation tool have been developed to properly compare treatment arms for decision making, while controlling Type I and Type II error rates. Some examples of treatment arm comparisons are available using data from oncology studies.

Conclusion We propose a practical approach to consider the efficacy-safety tradeoff and efficiently compare treatments based on categorical outcomes. The joint probit model considers the correlation between efficacy and toxicity to support multivariate decision-making and efficiently determines whether a treatment is clinically superior to another, by reducing the multidimensional outcome to a single mean utility score. In addition, the benefit-risk ratio is often considered to compare multiple dose levels, looking for the optimal dose. The proposed utility score is useful in summarizing the benefit-risk ratio in early drug development. The statistical test we propose can also be used for dose optimization or seamless designs and combined with commonly used study designs, such as Group Sequential Design.

38: Hierarchical Composite Endpoints and Win Ratio Methods in Cardiovascular Trials: a Systematic Review and Consequent Guidance

Ruth Owen^{1,2,3}, John Gregson¹, Dylan Taylor^{2,3}, David Cohen^{4,5}, Stuart Pocock¹

¹London School of Hygiene and Tropical Medicine, United Kingdom

²Centro Nacional de Investigaciones Cardiovasculares, Spain

³Oxon Epidemiology, Spain

⁴Cardiovascular Research Foundation, NY USA

⁵St. Francis Hospital, NY USA

Introduction The value of hierarchical composite endpoints (and their analysis using the win ratio) is being increasingly recognised, especially in cardiology trials. Their reporting in journal publications has not been previously explored.

Methods A search of 14 general medical and cardiology journals was done using 13 search terms including “hierarchical composite”, “win ratio”, and “Finkelstein Schoenfeld” during 01/Jan/2022 to 31/Jan/2024. We identified 61 articles (from 36 unique trials) that included

analyses using the win ratio. For multiple such articles from the same trial, we selected the most major (or first) one. A standardized proforma was completed by two reviewers (DT+RO), with any inconsistencies resolved by consensus.

Results Of the 36 trials identified, 10 were in NEJM, 20 were primary publications, and 10 had win ratio as the primary analysis. Most (N=26) were drug trials, but trials of device/surgery (N=7) and treatment strategies (N=3) also occurred. The most common conditions were heart failure (N=15) and ischemic heart disease (N=5).

The choice of hierarchical components varied: nearly all trials (N=32) had mortality as the first comparison, 30 of which had non-fatal events next. The number of non-fatal event components ranged from 0 (4 trials) to 6 (2 trials). In 27 trials, at least one component was a quantitative outcome, most commonly a quality-of-life score, of which 12 defined a minimal margin to claim a win/loss. Hierarchies ranged from 1 to 9 components, with 3 (N=11) and 4 (N=6) components being most common.

Trials usually reported the unmatched win ratio, its 95% CI and Finkelstein-Schoenfeld p-value, with results commonly presented using flowcharts (N=10) or bar charts (N=12). Win odds (4 trials) and win difference (3 trials) were occasionally reported. Stratified (9 trials) and covariate-adjusted analyses (1 trial) were not common. Of the 28 trials that reported the percentage of tied comparisons, 8 had <10% ties whilst 5 had >70% ties.

Specific examples will be presented to illustrate the diversity of good (and sometimes bad) practice in the use and reporting of the win ratio. We conclude with a set of recommendations for future use.

Discussion This systematic review is the first to document the diversity of uses of hierarchical composite endpoints and win ratio analyses in journal publications. This portfolio of mostly appropriate applications in cardiovascular trials suggests that hierarchical composite outcomes could be relevant in other diseases where treatment response cannot be captured by a single endpoint.

39: Power Calculation using the Win-Ratio for Composite Outcomes in Randomized Trials

David Kronthaler¹, Felix Beuschlein², Sven Gruber², Matthias Schwenkglenks³, Ulrike Held¹

¹Epidemiology, Biostatistics and Prevention Institute, Department of Biostatistics, University of Zurich, Switzerland

²Department of Endocrinology, Diabetology and Clinical Nutrition, University Hospital Zurich, University of Zurich, Switzerland

³Health Economics Facility, Department of Public Health, University of Basel, Switzerland

Background The use of composite outcomes is common in clinical research. These can include, for example, death from any cause and any untoward hospitalization, and corresponding effect measures would be the risk ratio or the hazard ratio typically addressing the time to first occurrence of any of the two events. In these situations, the hierarchy of the outcomes is ignored, and the combination of different outcome distributions is difficult.

Methods We used the win-ratio approach (Pocock et al. 2024) for the design and sample size calculation of a randomized controlled trial in patients suspected for primary aldosteronism. The win-ratio assumes N_T and N_C patients in treatment and control group, resulting in $N_T \times N_C$ pairwise comparisons of patients in treatment and control group. The win-ratio is then calculated as $R_W = N_W/N_L$, with N_W and N_L being the counts of wins and losses of patients in the treatment group.

The trial has a composite outcome with the following hierarchy:

I Elevated blood pressure (binary, according to WHO definition) and

II Defined daily dose (DDD) of blood pressure medication.

To assess for each comparison whether the patient of the treatment group is the winner or the loser, first hierarchy I, and upon a tie, hierarchy II outcomes are compared. As reference, the power of the trial was compared to a standard sample size calculation for a binary and a continuous outcome with the same specifications.

Results The power of the trial was assessed with 1000 simulation runs, and with $N_T = N_C = 300$ patients, assuming 15% drop-out. Our simulation showed that the resulting power of the trial was 85% and the estimated win-ratio R_W was 1.3. Under identical assumptions, standard power calculation would have resulted in 30% power for the hierarchy I outcome, and 73% power for the hierarchy II outcome.

Conclusion While the win-ratio has been employed in secondary analyses of randomized trials, it has rarely been used at study design level. Sample size calculation using the win-ratio as effect measure is efficient from a methodological perspective, and it captures well the complexities of using potentially censored composite outcomes with a hierarchy in clinical research.

References Pocock, Stuart J, John Gregson, Timothy J Collier, Joao Pedro Ferreira, and Gregg W Stone. 2024. "The Win Ratio in Cardiology Trials: Lessons Learnt, New Developments, and Wise Future Use." *European Heart Journal* 45 (44): 4684–99. <https://doi.org/10.1093/euheartj/ehae647>.

40: Feasibility of Propensity Score Weighted Analysis in Rare Disease Trials: a Simulation Study

Alexander Przybylski¹, Francesco Ambrosetti², Lisa Hampson², Nicolas Ballarini²

¹Novartis, UK

²Novartis, Switzerland

Introduction Clinical trials in rare diseases often face challenges due to small sample sizes and single-arm non-randomized designs, which increase the risk of confounding bias. Propensity scoring (PS) methods are commonly applied to mitigate such biases. However, in small samples, the ability to fit adequate PS models that reduce covariate imbalance has not been widely studied. In the context of an anticipated large treatment effect where the response probability on control is very low, the statistical challenges of using PS weighting for treatment effect estimation are further complicated. Our aim was to evaluate the feasibility and performance of PS methods under these specific conditions.

Methods A simulation study was conducted to assess the impact of covariate imbalance, sample size, and treatment effect size on the feasibility and performance of several estimators and intervals for the average treatment effect in the treated (ATT; expressed as a difference in marginal risks). The focus was on two key baseline covariates; large treatment effects informed by prior knowledge; and a small sample size of 15 subjects per arm. Weighted and unweighted ratio estimators, a hybrid approach incorporating PS model convergence and covariate imbalance criteria, and standardization-based estimators were evaluated according to estimator convergence rate, the probability of proceeding with an indirect comparison based on measures of imbalance (standardized mean difference; SMD), and bias. Coverage probabilities of intervals were also calculated.

Results Standardization-based estimators were unreliable due to low sample size and complete separation issues. Propensity score models could be estimated and were able to reduce imbalance even with small sample sizes and high imbalance. Setting a 0.1 SMD threshold for adequate covariate balance, 25% of simulation runs met the criteria for performing the indirect comparison analysis. The use of a less conservative 0.25 threshold for SMD increased this probability to 50% while maintaining acceptable bias and coverage probability. Condi-

tional on observing at least one response in the control arm, average conditional bias was marginally improved via propensity score weighting.

Conclusions Propensity score weighting methods can address confounding biases in non-randomized studies, even with small sample sizes and large treatment effects. However, in our setting, the most suitable approach involved using a hybrid method that combines pre-specified criteria for performing the indirect comparison.

41: A Basket Trial for Rare Diseases, with a Crossover Design for Its Substudies: a Simulation Study

Elena G Lara, Steven Teerenstra, Kit C.B. Roes, Joanna IntHout

Radboud University Medical Center, Netherlands, The

Background Recent advancements in precision medicine generate therapy options for rare diseases. Assessing a new treatment targeted to a rare disease subgroup can make recruiting the required sample size even more challenging. Current work recommends grouping rare diseases in a basket trial, where one drug is evaluated in multiple diseases based on a shared etiology (e.g. gene mutation). This allows to include more patients and to borrow information between substudies. A further recommendation to improve efficiency for trials involving chronic and stable conditions is the use of crossover designs. Our research focuses on basket trials with a crossover design for the substudies. These may increase precision of the estimated treatment effect, both by borrowing information across substudies, as well as efficient substudy design.

Methods In this study, we evaluated the operating characteristics of basket trials where each substudy corresponds to a crossover design via Monte Carlo simulation. We generated realistic scenarios related to the SIMPATHIC project, under parallel and crossover designs, and with different numbers of substudies (from 2 to 9). We applied estimation methods including random-effects meta-analysis, Bayesian hierarchical modelling (BHA), EXNEX, adaptive lasso, stratified analysis and naïve pooling. And we studied the bias, precision, power and false positive rate of the substudy estimates as well as the trial overall estimates.

Results The efficiency gains of crossover designs in conventional trials are also present in basket trials. Methods that use information borrowing improve estimation of substudy treatment effects in terms of increased precision. This increase in precision is lower in substudies with a crossover design compared to the parallel-group design with the same number of patients; borrowing in this setting also results in lower shrinkage. Among the borrowing

methods evaluated, EXNEX seems the most able to discriminate between substudies with a true effect and those with a small or null effect. Meta-analysis, BHA and naïve pooling achieve the highest power for the overall estimate, although this power is low when the treatment had a true effect in less than half of the substudies.

Conclusion The incorporation of crossover designs to basket trial substudies - when assumptions are met - results in a more efficient design and practicable sample sizes compared to parallel-group designs. Besides, adding randomization and a control per substudy provides a more valid inference than a single arm design. Altogether, this design can facilitate drug development for rare diseases.

Project funded by Horizon Europe (Grant no. 101080249).

42: Comparing Randomized Trial Designs in Rare Diseases with Longitudinal Models: a Simulation Study Showcased by Autosomal Recessive Cerebellar Ataxias

Niels Hendrickx¹, France Mentré¹, Alzahra Hamdan², Mats Karlsson², Andrew Hooker², Andreas Traschütz^{3,4}, Cynthia Gagnon⁵, Rebecca Schüle⁶, ARCA Study group⁷, EVIDENCE-RND Consortium⁷, Matthis Synofzik^{3,4}, Emmanuelle Comets^{1,8}

¹Université Paris Cité, IAME, Inserm, F-75018, Paris, France

²Pharmacometrics Research Group, Department of Pharmacy, Uppsala University, Uppsala, Sweden

³Division Translational Genomics of Neurodegenerative Diseases, Hertie Institute for Clinical Brain Research (HIH), University of Tübingen, Tübingen, Germany

⁴German Center for Neurodegenerative Diseases (DZNE), Tübingen, Germany.

⁵Centre de Recherche du CHUS Et du Centre de Santé Et Des Services Sociaux du Saguenay-Lac-St-Jean, Faculté de Médecine, Université de Sherbrooke, Québec, Canada.

⁶Hertie-Center for Neurology, University of Tübingen, Tübingen, Germany

⁷Group author

⁸Univ Rennes, Inserm, EHESP, Irset - UMR_S 1085, 35000, Rennes, France.

Background Parallel designs with an end-of-treatment analysis are commonly used for randomised trials (1), but they remain challenging to conduct in rare diseases due to small sample size and heterogeneity. A more powerful alternative could be to use model-based approaches (2,3). We investigated the performance of longitudinal modelling to evaluate disease-modifying treatments in rare diseases using simulations. Our setting was based on a model describing the progression of the standard clinician-reported outcome SARA score in

patients with ARCA (Autosomal Recessive Cerebellar Ataxia), a group of ultra-rare, genetically defined, neurodegenerative diseases (4).

Methods We performed a simulation study to evaluate the influence of trials settings on their ability to detect a treatment effect slowing disease progression, using a previously published non-linear mixed effect logistic model (5). We compared the power of parallel, crossover and delayed start designs (6,7), investigating several trial settings: trial duration (2 or 5 years); disease progression rate (slower or faster); magnitude of residual error ($\sigma=2$ or $\sigma=0.5$); number of patients (100 or 40); method of statistical analysis (longitudinal analysis with non-linear or linear models; standard statistical analysis), and we investigated their influence on the type 1 error and corrected power of randomised trials.

Results In all settings, using non-linear mixed effect models resulted in controlled type 1 error and higher power (88% for a parallel design) than a rich (75% for a parallel design) or sparse (49% for a parallel design) linear mixed effect model or standard statistical analysis (36% for a parallel design). Parallel and delayed start designs performed better than crossover designs. With slow disease progression and high residual error, longer durations are needed for power to be greater than 80%, 5 years for slower progression and 2 years for faster progression ataxias.

Conclusion In our settings, using non-linear mixed effect modelling allowed all three designs to have more power than a standard end-of-treatment analysis. Our analysis also showed that delayed start designs are promising as, in this context, they are as powerful as parallel designs, but with the advantage that all patients are treated within this design.

- References**
- (1) E9 Statistical Principles for Clinical Trials, <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/e9-statistical-principles-clinical-trials>, 2020
 - (2) Synofzik et al. Neuron 2019
 - (3) Buatois et al; Statistics in Medicine 2021
 - (4) Karlsson et al. CPT Pharmacometrics Syst Pharmacol 2013
 - (5) Hamdan et al. CPT 2024
 - (6) Liu-Seigfert et al. PLoS ONE 2015
 - (7) Wang et al. Pharmaceutical Statistics 2019

43: Sequential Decision Making in Basket Trials Leveraging External-Trial Data: With Applications to Rare-Disease Trials

Giulia Risca¹, Stefania Galimberti¹, Maria Grazia Valsecchi¹, Haiyan Zheng²

¹Bicocca Bioinformatics Biostatistics and Bioimaging B4 Center, Department of Medicine and Surgery, University of Milan-Bicocca, Monza, Italy

²Department of Mathematical Sciences, University of Bath, Bath, UK

Introduction Rare diseases present unique challenges in the design of clinical trials due to a small pool of eligible patients. Planning rare-disease studies within a basket trials, which can simultaneously evaluate a new treatment in patients with a shared disease trait, is practical because of borrowing strength from relevant patient subgroups. Motivated by a real rare-disease trial under planning, we develop a Bayesian sequential design that allows incorporation of both external-trial and within-trial data for basket trials involving rare diseases.

Methods We consider two subgroups of patients to receive the same treatment before deciding on if a third one would be treated in the basket trial. The EXNEX method¹ is extended to include a prior mixture component formed using external-trial data. That is, the treatment effects in those three subgroups are assumed to be exchangeable, or non-exchangeable but consistent with the external-trial data, or completely extreme. On the completion of the first two subgroups, our Bayesian meta-analytic-predictive model is used to obtain the predictive probability (PP) of an efficacious treatment in the third subgroup. Interim futility assessment is guided using a power spending function.

Results We assess the performance of this design through simulations, which results sensitive to the choice of certain parameters (e.g., prior mixture weights, cut-offs for the interim and the final analyses). Specifically, the PPs at the first interim are highly dependent on the different allocation weights. Pessimistic scenarios have large variability in PPs depending on whether the exchangeability or the prior-data consistency assumption is violated. However, it is generally robust when there is strong belief in a highly effective treatment and all models seem to accurately estimate the true treatment effect in each subgroup in terms of bias and mean squared error. Finally, the marginal type I error is always well controlled.

Conclusions In conclusion, our method allows mid-course adaptation and ethical decision-making. It is novel and can address critical gaps in rare diseases. The principles are generalizable to other contexts.

References 1. Neuenschwander, B., Wandel, S., Roychoudhury, S. & Bailey, S. (2016) Robust exchangeability designs for early phase clinical trials with multiple strata. *Pharmaceutical Statistics*, 15, 123–134. Available from: <https://doi.org/10.1002/pst.1730>

44: Adaptive Designs and Bayesian Approaches: The Future of Clinical Trials

Anjali Yadav

JSS Medical Research Asia Pacific Pvt. Ltd., India

Background / Introduction Traditional clinical trial designs rely on fixed protocols that do not allow for modifications once the study is initiated. This rigidity can lead to inefficiencies, ethical concerns, and prolonged development timelines. Adaptive designs provide a flexible framework that permits pre-specified modifications based on interim analyses, improving resource allocation and patient outcomes. Meanwhile, Bayesian approaches leverage prior knowledge and continuously update probabilities, offering a more dynamic and intuitive method for decision-making. The integration of these methodologies has the potential to revolutionize clinical trial efficiency, particularly in the era of precision medicine and rare disease research.

Methods This study reviews key adaptive design strategies, including group sequential, response-adaptive, and platform trials, highlighting their statistical foundations and regulatory considerations. Bayesian methodologies, such as Bayesian hierarchical modeling and predictive probability monitoring, are explored in the context of trial adaptation and decision-making. Case studies from oncology, vaccine development, and rare disease trials are examined to illustrate the real-world application and advantages of these approaches.

Results Adaptive designs have demonstrated significant reductions in trial duration and costs while maintaining scientific integrity. Bayesian methods have enhanced decision-making by incorporating historical data and real-time learning, leading to more efficient dose-finding, early stopping for efficacy or futility, and improved patient allocation. Regulatory agencies, including the FDA and EMA, have increasingly supported these innovative methodologies, providing frameworks for their implementation. Case studies highlight improved success rates, patient safety, and ethical advantages compared to traditional approaches.

Conclusion The adoption of adaptive designs and Bayesian approaches is transforming clinical research by making trials more efficient, ethical, and informative. While challenges remain, including regulatory acceptance, operational complexity, and computational demands, ongoing advancements in statistical methods and trial simulations continue to enhance their feasibility. The future of clinical trials lies in the strategic integration of these methodologies, fostering a more flexible and patient-centric approach to drug development.

Tuesday Posters at ETH

Tuesday, 2025-08-26 09:15 - 10:45, ETH, UG hall

1: Leveraging Wearable Data for Probabilistic Imputation in Cardiovascular Risk Calculators.

Antoine Faul¹, Patric Wyss², Anja Mühlemann¹, Manuela Moraru³, Danielle Bower³, Petra Stute³, Ben Spycher², David Ginsbourger¹

¹Institute of Mathematical Statistics and Actuarial Science, University of Bern, Bern, Switzerland

²Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland

³Department of Obstetrics and Gynecology, University Women's Hospital, Bern, Switzerland

Wearable technology for health data collection is rapidly expanding enabling continuous, non-invasive monitoring of physiological parameters such as heart rate variability and physical activity. This advancement offers promising improvements for cardiovascular disease (CVD) risk prediction, which traditionally depends on clinical measurements often requiring time-consuming and costly healthcare visits.

This study analyzes data from 193 female participants, aged 40 to 69, gathered during an observational study at Inselspital, Bern. Participants provided comprehensive medical and personal information, supplemented by wearable data collected with Garmin Vivosmart 3 devices over a week.

In this work, we explore the potential of replacing systematically missing inputs with probabilistic predictions derived from wearable data and self-reported information. By integrating this uncertainty into a risk calculator, we aim to provide probabilistic assessments of cardiovascular risk. Our approach uses an interpretable statistical model based on Gaussian copulas. This method flexibly characterizes the joint distribution, employing distinct marginals and a Gaussian dependence structure to facilitate analytical conditioning.

We extend the approach outlined by Mühlemann [1] by addressing the challenge of high dimensionality of smartwatch data. For this we focus on selected features obtained by both supervised and unsupervised dimensionality reduction techniques.

Proper scoring rules, such as the CRPS and the Brier Score, are employed to assess the

quality of probabilistic predictions. We also compare various methods by cross validation in the context of high vs low CVD risk classification.

Our results demonstrate that wearable data can help in substituting clinical missing inputs in cardiovascular risk calculators, provided that an efficient dimension reduction step is implemented. However, the gains in predictive performance are moderate, suggesting that further exploration of advanced dimensionality reduction techniques could be beneficial.

[1] MÜHLEMANN, Anja, STANGE, Philip, FAUL, Antoine, et al. Comparing imputation approaches to handle systematically missing inputs in risk calculators. PLOS Digital Health, 2025, vol. 4, no 1, p. e0000712.

2: Stepwise Prediction of Tuberculosis Treatment Outcomes Using XGBoost and Feature-Level Analysis: A Multi-Stage Approach to Clinical Decision Support

Linfeng Wang, Jody Phelan, Taane Clark

London School of Hygiene and Tropical Medicine, United Kingdom

Tuberculosis (TB) remains a global health crisis, with multidrug-resistant (MDR-TB) and extensively drug-resistant (XDR-TB) strains posing significant challenges to treatment. Utilizing the extensive TB Portals database, comprising clinical, radiological, demographic, and genomic data from 15,997 patients across high-burden countries, we developed an XGBoost-based machine learning model to predict treatment outcomes. Our approach categorizes features into four diagnostic evidence data categories: demographic, microbiology and disease state, x-ray result variables and treatment variables. This framework enables the model to progressively incorporate available data while maintaining robust predictive performance, even in the presence of missing values typical of real-world healthcare settings. The model achieved high predictive accuracy (AUC-ROC: 0.96, F1-score: 0.94), with key predictors including age of onset, drug resistance, and treatment adherence. Regional analysis highlighted variability in performance, underscoring the potential for localized model adaptation. By accommodating missing data at various diagnostic stages, our model provides actionable insights for personalized TB treatment strategies and supports clinical decision-making in diverse and resource-constrained contexts.

3: The Use of Variable Selection in Clinical Prediction Modelling for Binary Outcomes: a Systematic Review

Xinrui Su, Gareth Ambler, Nathan Green, Menelaos Pavlou

Department of Statistical Science, University College London

Background Clinical prediction models can serve as important tools, assisting in medical decision-making. Concise, accurate and interpretable models are more likely to be used in practice and hence an appropriate selection of predictor variables is viewed as essential. While many statistical methods for variable selection are available, data-driven selection of predictors has been criticised. For example, the use of variable selection with very low significance levels can lead to the exclusion of variables that may improve predictive ability. Hence, their use has been discouraged in prediction modelling. Instead, selection of predictors based on the literature and expert opinion is often recommended. Recent sample size guidelines also assume that predictors have been pre-specified, and no variable selection is performed. This systematic review aims to investigate current practice with respect to variable selection when developing clinical prediction models using logistic regression.

Methods We focused on published articles in PubMed between 1-21 October 2024 that developed logistic prediction models for binary health outcomes. We extracted information on study characteristics and methodology.

Results In total 141 papers were included in the review. We found that almost all papers (140/141) used variable selection. Univariable selection (UVS) was by far the most commonly reported method; it was used solely or sequentially alongside other methods in 78% (110/141) of papers. It was followed by backwards elimination (BE) (60/141, 43%), 'with bulk removal' (BR) from a single model (58/141, 41%) and LASSO (35/141, 25%). UVS and BE were frequently applied together (45/139, 32%), as were UVS and BR (43/139, 31%).

Conclusions Despite criticisms regarding the uncritical use of data-driven variable selection methods, surprisingly almost all studies in this review employed at least one such method to reduce the number of predictors, with many studies using multiple methods. Traditional methods such as UVS and BE, as well as more modern techniques such as LASSO, are still commonly used. In the pursuit of parsimonious, as well as accurate risk models, model developers must be cautious when using methods based on significance testing, particularly with very low significance levels. However, methods such as LASSO, which directly aim to optimise out-of-sample predictive performance while also removing redundant predictors, may be promising and merit attention.

4: Comparison of Methods for Incorporating Related Data when Developing Clinical Prediction Models: A Simulation Study

Haya Elayan, Matthew Sperrin, Glen Martin, David Jenkins

University of Manchester, United Kingdom

Background Clinical Prediction Models (CPMs) are algorithms that compute an individual's risk of a diagnostic or prognostic outcome, given a set of their predictors. Guidance states CPMs should be constructed using data directly sampled from the target population. However, researchers might also have access to additional and potentially related datasets (ancillary data) originating from different time points, countries, or healthcare settings, which could support model development, especially when the target dataset is small.

A critical consideration in this context is the potential heterogeneity between the target and ancillary datasets due to data distribution shifts. These occur when the distributions of predictors, event rates, or the relationships between predictors and outcome differ. Such shifts can negatively affect CPMs performance in the target population. We aim to investigate in what situations and using which methods, the ancillary data should be incorporated when developing CPMs. Specifically, if the effectiveness of utilising the ancillary data is influenced by the heterogeneity between the available datasets, and their relative sample sizes.

Methods We conducted a simulation study to assess the impact of these factors on CPM performance when ancillary data is available. Target and ancillary populations were generated with varying degrees of heterogeneity. CPMs were developed using naive logistic regression (developed on data from target only), Logistic and Intercept regression updating methods (developed on ancillary data and updated to target), and importance weighting using propensity scores (developed on all available data, while weighting the ancillary data samples based on their similarity to the target). These models were then validated on independent data drawn from the same data-generating mechanism as the target population, using calibration, discrimination, and prediction stability metrics.

Results and Conclusion Incorporating ancillary data consistently improved performance compared to using the target data only, especially when the target sample size was small. Both Logistic and Intercept Recalibration improve performance over naïve regression in most scenarios. However, the former showed greater variability in calibration slopes and more instability in calibration curves, while the latter performed worse in calibration slope under predictor-outcome association shift.

Importance weighting using propensity scores showed consistent results with improved performance to other methods in many scenarios, particularly under predictor-outcome association

shift.

While this study investigates a-priori known data distribution shifts, their presence and type in practical settings are often unknown. Therefore, we recommend using importance weighting method for its robustness and stability across varied scenarios.

5: A Systematic Review of Methodological Research on Multi-State Prediction Models

Chantelle Cornett, Glen Martin, Alexander Pate, Victoria Palin

University of Manchester, United Kingdom

Background

Prediction models use information about a person to predict their risk of disease. Across health, patients transition between multiple states over time, such as health states or disease progression. Here, multi-state models are crucial, but these models require additional methodological considerations and their application in prediction modelling remains scarce. The methodological state-of-play of these methods in a prediction context has not been summarised.

Objectives

This systematic review aims to summarise and critically evaluate the methodological literature on multi-state models, with a focus on development and validation techniques.

Methods

A comprehensive search strategy was implemented across PubMed, Scopus, Web of Science, arXiv to identify methodological papers on multi-state models up to 7th October 2024. Papers were included if they focused on methodological innovation, such as sample size determination, calibration, or novel computational methods; we excluded purely applied papers. Methodological details were extracted and summarised using thematic analysis.

Results The search identified 14,788 papers. After the title and abstract screening, there were 443 papers for full-text screening, of which 299 papers were included.

Preliminary findings from these studies reveal the majority of methodological research falls into the following groups:

1. Techniques for estimating transition probabilities, state occupation time, and hazards.

2. Hypothesis testing.
3. Variable selection techniques.

This presentation will overview the themes of methodological work, the limitations/gaps in methodological literature in this space, and outline areas for future work.

Conclusions

Early results highlight progress in the methodological development of multi-state models and emphasise areas requiring further attention, such as more research into sample size and robust validation practices. The final results of this study aim to guide future research and support the adoption of best practices in the use of multi-state models.

6: Assessing the Robustness of Prediction Models: A Case Study on in-Hospital Mortality Prediction using MIMIC-III and MIMIC-IV

Alan Balendran¹, Raphaël Porcher^{1,2}

¹Université Paris Cité, Université Sorbonne Paris Nord, INSERM, INRAE, Centre for Research in Epidemiology and StatisticS (CRESS)

²Centre d'Épidémiologie Clinique, Assistance Publique-Hôpitaux de Paris, Hôtel-Dieu

Clinical prediction models have become increasingly prevalent due to the availability of large healthcare datasets. While these models often achieve strong predictive performance, their *robustness*—their ability to remain stable under various perturbations—remains underexplored. However, models may experience significant performance degradation when tested on perturbed data (e.g., noisy data or datasets collected at different time points). Understanding how robust a prediction model is is essential for ensuring reliable clinical decision-making.

Building on an existing framework that identified eight key robustness concepts in healthcare (*Balendran, A., Beji, C., Bouvier, F. et al. A scoping review of robustness concepts for machine learning in healthcare. npj Digit. Med. 8, 38 (2025)*), we evaluate the robustness of different machine learning models using real-world critical care data from intensive care unit (ICU) patients.

We utilise the MIMIC-III and MIMIC-IV critical care databases to predict in-hospital mortality based on patient data from their first 24 hours of ICU admission. The dataset includes vital signs, laboratory test results, and demographic information (*Johnson, A., Pollard, T., Shen, L. et al. MIMIC-III, a freely accessible critical care database. Sci Data 3, 160035 (2016)*).

To develop prediction models, we explore a range of machine learning approaches, from linear models such as logistic regression and LASSO to more complex tree-based methods, including random forest and gradient boosting. Additionally, we assess deep learning models, including a multilayer perceptron (MLP) and the recently introduced transformer-based model TabPFN (*Hollmann, N., Müller, S., Purucker, L. et al. Accurate predictions on small data with a tabular foundation model. Nature 637, 319–326 (2025)*), which has been reported to outperform traditional gradient boosting techniques.

Each model is evaluated across multiple robustness concepts, including input perturbations, label noise, class imbalance, missing data, temporal validation, and subgroup analysis. To better reflect real-world clinical settings, we introduce varying levels of noise and test different scenarios for some concepts.

Our findings demonstrate that no model is consistently robust across all concepts, with some models being particularly sensitive to specific perturbations. Our result highlights that relying solely on standard performance metrics within a dataset does not account for potential deviations that can be encountered in real clinical settings. We advocate for robustness assessments as a crucial component of model evaluation and selection in healthcare.

7: The Influence of Variable Selection Approaches on Prediction Model Stability in Low-Dimensional Data: From Traditional Stepwise Selection to Regularisation Techniques

Noraworn Jirattikanwong¹, Phichayut Phinyo¹, Pakpoom Wongyikul¹, Natthanaphop Isaradech², Wachiranun Sirikul², Wuttipat Kiratipaisarl²

¹Department of Biomedical Informatics and Clinical Epidemiology (BioCE), Faculty of Medicine, Chiang Mai University, Chiang Mai, Thailand

²Department of Community Medicine, Faculty of Medicine, Chiang Mai University, Chiang Mai, Thailand

Introduction Prediction model instability presents significant challenges to clinical decision-making and may lead to patient harm. While several factors, such as dataset size and algorithm choice, are known to affect stability, evidence on how specific modelling decisions, particularly variable selection methods, influence stability remains limited. This study examines the impact of different variable selection approaches on prediction stability and model performance.

Methods The German HOPE dataset of 9,924 patients, previously used to develop an anxiety

prediction model was used. We generated three datasets of different sizes (0.5, 1, and 2 times the base size), where the base size was determined using Riley's minimum sufficient sample size method. We defined 61 candidate parameters and replicated the model to predict anxiety using logistic regression. Six variable selection approaches were examined: (1) UNIVAR – univariate screening followed by backward elimination, (2) FULL – full model including all variables, (3) FORWARD – forward selection, (4) BACKWARD – backward elimination, (5) LASSO – least absolute shrinkage and selection operator, and (6) ELASTIC – elastic net. Model performance was evaluated in terms of discrimination and calibration. Optimism in performance metrics and mean absolute prediction error (MAPE) were estimated using the bootstrap internal validation procedure proposed by Riley and Collins.

Results All variable selection approaches exhibited a similar level of discrimination at the base size and twice the base size. In contrast, at half the base size, both discrimination and calibration measures varied considerably. FULL achieved the highest discrimination in the smallest dataset but consistently displayed poor calibration across all sample sizes. Regularisation approaches (i.e., LASSO and ELASTIC) were well-calibrated across all dataset sizes, whereas traditional stepwise selection methods (i.e., UNIVAR, FORWARD, and BACKWARD) were only well-calibrated when the sample size was twice the base size. In terms of stability, both regularisation approaches had lower MAPE than others at the base size and twice the base size, while FULL showed lower MAPE at half the base size. All approaches required at least twice the minimum sufficient sample size to achieve a high level of individual stability.

Conclusion Variable selection using regularisation is recommended, provided the sample size is sufficiently large. When sample sizes are around half the base size, regularisation approaches may still outperform other techniques in terms of stability and calibration. While FULL resulted in a modest improvement in stability, it exhibited significantly poorer calibration compared to UNIVAR, FORWARD, and BACKWARD.

8: Early-Detection of High-Risk Patient Profiles Admitted to Hospital with Respiratory Infections using a Multistate Model

João Pedro Carmezim¹, Cristian Tebé¹, Natàlia Pallarès¹, Roger Paredes¹, Cavan Reilly²

¹Germans Trias i Pujol Research Instituteand Hospital (IGTP), Spain

²University of Minnesota

Background This study aims to identify clinically relevant prognostic factors associated with oxygen support, death or hospital discharge in a global cohort of adult patients with Influenza

or COVID-19 using a multistate model.

Methods Data was drawn from a cohort of adult patients diagnosed with respiratory infections admitted to a hospital of the Strategies and Treatments for Respiratory Infections and Viral Emergencies (STRIVE) research group. The study evaluates socio-demographic factors, medical history, comorbidities, vaccination status, virus type and clinical symptoms as prognostic factors. The multistate model was defined with the following states: hospital admission, noninvasive ventilation, invasive ventilation, oxygen support discharge, hospital discharge and death. The model estimates cause-specific hazard ratios, cumulative hazards and transition probabilities.

Results A total of 4968 patients were included where the median age was 62.1 and the percentage of females was 47.9%. The number of patients that needed noninvasive ventilation was 1906 (38.4%), 277 (5.6%) required invasive ventilation , and 275 (5.5%) died. Demographic and clinical risk profiles revealed distinct progression pathways, and visualization using trajectory plots highlighted how risk factors influenced movement through disease states.

Conclusion This study highlights the utility of a multistate model in mapping the progression of respiratory infections, providing critical insights into high-risk patient profiles. Transition probability trajectories provide clinicians with data to predict outcomes and, ideally, could help to plan resource allocation for these patients.

9: Investigating Fair Data Acquisition for Risk Prediction in Resource-Constrained Settings

Ioanna Thoma¹, Matthew Sperrin², Karla Diaz Ordaz³, Ricardo Silva³, Brieuc Lehmann³

¹The Alan Turing Institute, London, United Kingdom

²Division of Informatics, Imaging & Data Sciences, The University of Manchester, Manchester, United Kingdom

³Department of Statistical Science, University College London, London, United Kingdom

Introduction Accurate risk prediction relies on robust clinical prediction models (CPMs), yet their reliability, generalisability, and fairness can be constrained by the available data. While additional covariates may improve risk prediction, collecting them for an entire population might not always be feasible due to resource constraints. For example, genetic testing can provide additional predictive power when combined with a clinical risk model, but a

population-wide rollout may not be financially viable. A key question is how to allocate resources, prioritising individuals for whom additional (genetic) testing would benefit most. This framework optimises utility and fairness when choosing between a baseline prediction model and a more costly but potentially more informative augmented model.

Methods We develop a framework that quantifies the potential benefit to fairness and accuracy of a CPM when assessing policies for acquiring additional information for a subset of individuals. A specific use case is deploying an integrated tool that combines a traditional CPM, based on clinical risk factors, with a polygenic risk score (PRS). The goal is to evaluate the utility gained from such data integration. This involves comparing the outcomes of a conventional CPM with those of an integrated tool to assess how risk categorisation shifts when genetic information is incorporated.

Results We apply our methodology to cardiovascular disease (CVD) risk prediction on a UK Biobank cohort of 96884 individuals aged 40-75. Transitions in risk classification help identify populations that benefit most from genetic score integration. Once these population subgroups have been identified, we define sub-sampling policies to determine which individuals should be selected based on their covariates and existing model uncertainty. We investigate deterministic and stochastic policies that also account for varying subgroup proportions, ensuring a representative and fair sample composition. The methodology identifies age and gender groups that experience the most significant shifts in risk classification when transitioning from the baseline to the integrated model.

Conclusion This framework has the potential to guide future data collection strategies, helping to prioritise population subgroups that need it the most. While our application focuses on the evaluation of an integrated tool for CVD risk prediction, we expect the methodology to be broadly applicable and can be adapted to a variety of predictive models across the disease spectrum.

10: A Critical Benchmark of Bayesian Shrinkage Estimation for Subgroup Analysis

Sebastian Weber¹, Björn Bornkamp¹, David Ohlssen²

¹Novartis Pharma AG, Switzerland

²Novartis Pharmaceuticals, USA

The estimation of subgroup specific treatment effects is known to be a statistically diffi-

cult problem. We suggest to evaluate different estimation approaches using a benchmark. This benchmark is based on scoring the predictive distribution for the subgroup treatment effect using late phase clinical trial data comprising normal, binary and time-to-event endpoints. Bayesian shrinkage estimation models for subgroups are traditionally applied to non-overlapping subgroups using hierarchical models. This implies that several models need to be fitted to the same data set when several subgroup defining variables are of interest. Recently Wolbers et al (2024) propose to use a single global regression model using priors such as horseshoe priors to induce shrinkage for the used model. This method has the benefit that there is no need to create a disjoint space of subgroups. Thus, overlapping subgroups can be investigated with a single model avoiding the need to refit a given data set multiple times. We will compare the performance of different shrinkage approaches based on a real data benchmark. The evaluated approaches include no and full-shrinkage towards the overall treatment effect, Bayesian hierarchical shrinkage and more novel priors such as the global model prior R2D2 proposed by Zhang et al (2020).

11: Mathematical Modelling of Oxygenation Dynamics Using High-Resolution Perfusion Data: An Advanced Statistical Framework for Understanding Oxygen Metabolism

Mansour Taghavi Azar Sharabiani¹, Alireza Mahani², Richard Issitt³, Yadav Srinivasan⁴, Alex Bottle¹, Serban Stoica⁵

¹School of Public Health, Imperial College London, United Kingdom

²Statman Solution Ltd, United Kingdom

³Perfusion Department, Great Ormond Street Hospital for Children, London, United Kingdom

⁴Cardiac Surgery Department, Great Ormond Street Hospital for Children, London, United Kingdom

⁵Cardiac Surgery Department, Bristol Royal Children's Hospital, Bristol, United Kingdom

Background

Balancing oxygen supply and demand during cardiopulmonary bypass (CPB) is crucial to minimising adverse outcomes. Oxygen supply is determined by cardiac index (CI), haemoglobin concentration (Hb), and arterial oxygen saturation (SaO_2), whereas oxygen demand is driven by metabolism, which itself depends on body temperature (Temp). Actual oxygen consumption is driven by oxygen extraction ratio (OER), dynamically adapting to changes in oxygen supply and demand, yet the mechanisms of this adaptation remain poorly understood. We developed GARIX and eGARIX, mathematically extending classical time-series models to incorporate nonlinear dependencies, patient-specific variabilities and minute-by-minute OER dynamics.

Methods

GARIX is a time-series model that integrates exogenous variables (Cl, Hb, SaO₂, Temp) with a disequilibrium term representing the imbalance between oxygen consumption and temperature-dependent oxygen demand, initially modelled via a constant Q₀ framework (van't Hoff model). The model was trained on intraoperative data from 343 CPB operations (20,000 minutes) in 334 paediatric patients at a UK centre (2019–2021). eGARIX extends GARIX by relaxing the assumption of constant Q₀, introducing nonparametric temperature dependence (splines) and incorporating age, weight, and their interaction. Subgroup analyses explored OER responses across different age groups.

Results

GARIX identified that OER adapts in a two-phase process: a rapid adjustment phase (<10 minutes) and a slower phase lasting several hours. Equilibrium analysis estimated Q₀ = 2.25, indicating that oxygen demand doubles with every 8.5°C temperature increase. eGARIX demonstrated indexed oxygen demand following a nonlinear trajectory with age and weight, peaking at 3 years of age. In neonates and infants, oxygen demand correlated positively with weight, whereas in adolescents, the correlation was negative. Additionally, temperature dependence deviated from the classical Q₀ assumption, showing low sensitivity at mild hypothermia and high sensitivity at deep hypothermia. Younger patients exhibited a diminished OER response to Hb changes compared to older children.

Conclusions

Proposed GARIX and eGARIX represent mathematical extensions of classical time-series modelling, enabling a data-driven approach to studying oxygen metabolism during CPB. By harnessing vast amounts of recently available high-resolution perfusion data, these models compensate for the ethical limitations of direct human experimentation, providing a powerful framework to refine intraoperative oxygenation strategies. Our findings highlight the importance of advanced mathematical modelling in optimising personalised oxygen delivery strategies, adapting to individual patient characteristics, and enhancing our understanding of oxygen metabolism in paediatric CPB.

12: Marginal Structural Cox Model with Weighted Cumulative Exposure Modelling for the Estimation of Counterfactual Population Attributable Fractions

Yue Zhai¹, Ana-Maria Vilcu², Jacques Benichou^{2,3}, Lucas Morin², Agnès Fournier⁴, Anne Thiébaut², Vivian Viallon¹

¹Nutrition and Metabolism Branch, International Agency of Research in Cancer(IARC), Lyon, France

²High Dimensional Biostatistics for Drug Safety and Genomics Team, Université Paris-Saclay, UVSQ, Inserm, CESP, Villejuif, France

³Department of Biostatistics, Rouen University Hospital, Rouen, France

⁴Exposome and Heredity Team, CESP U1018, Université Paris-Saclay, UVSQ, Inserm, Gustave Roussy, Villejuif, France

Introduction Marginal structural Cox models (Cox MSMs) have become popular for estimating the causal effect of time-varying exposures on a time-to-event outcome, accounting for time-varying confounders affected by prior exposure levels. They can be combined with the weighted cumulative exposure (WCE) method to flexibly model the causal effect of past levels of the exposure on the hazard rate. This study evaluated the performance of the corresponding approach (Cox WCE MSM) based on regression B-splines, for the estimation of population attributable fractions (PAF) through extensive simulations.

Method Independent samples of 10,000 and 1,000 individuals, each with 100 regular visits of follow-up, were generated. In each sample, approximately 50% of individuals experienced the event of interest before the end of follow-up. For a given hazard ratio comparing “always exposed” to “never exposed”, we considered four scenarios, with different standardized weight functions reflecting how past exposure causally influences the current hazard rate as time elapses since exposure: i) monotonically decreasing weight; ii) bell-shaped weight; iii) constant weight; iv) current exposure only. Estimands of interest were the PAF and the causal effect function of past exposure. Various versions of Cox WCE MSMs were implemented to assess the influence of parameters like the number of knots and the length of time window. Additionally, we implemented two versions of Cox MSM accounting for only current exposure and unweighted cumulative exposure, respectively. The Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) were used for model selection.

Results PAF estimates produced by most Cox WCE MSMs were unbiased in scenarios i to iii, but were biased in scenario iv. The variance of Cox WCE MSMs was comparable to that of conventional Cox MSMs. Notably increasing the number of knots had little effect on variance. Models selected via either AIC or BIC provided unbiased PAF estimates across all scenarios. As for the causal effect of past exposure, although average estimates provided by Cox WCE MSMs were generally close to the true function, we observed large variation across samples, especially with smaller samples and weaker effects.

Conclusion Overall, Cox WCE MSMs selected by either AIC or BIC yielded unbiased estimates of counterfactual PAF. To ensure robust model selection, we recommend considering also the conventional Cox MSMs that account for current and unweighted cumulative exposure in the model selection process.

13: Lost in the Forest of Forest Plots? Practical Guidelines and an All-in-One Tool for Forest Plots

Hongqiu Gu, Yong Jiang, Hao Li

Beijing Tiantan Hospital, Capital Medical University, People's Republic of China

Background Forest plots are indispensable visualization tools in meta-analyses and other contexts of medical research, yet existing guidelines and implementation tools are often fragmented and lack a cohesive framework. In this study, we aimed to develop comprehensive guidelines and integrated tools to extend the applicability of forest plots across a wider range of research contexts.

Methods In consultation with a thorough review of existing literature and guidelines, combined with practical experience, we synthesized and developed a comprehensive classification system for forest plots driven by analysis methods. Additionally, we proposed key principles for their construction and created a versatile SAS macro to facilitate more effective application and communication of forest plots across various research scenarios.

Results We categorized forest plots into four main types that correspond to regression analysis, subgroup analysis, estimation analysis, and meta-analysis across 11 scenarios independent of study design. The five key principles for creating effective forest plots are providing comprehensive data, arranging items logically, ensuring accurate scaling, and applying aesthetic formatting. Furthermore, we developed versatile and integrated SAS tools that align with the framework and principles proposed.

Conclusion This guideline provides a versatile, integrated solution for applying forest plots across various research contexts. It is expected to lead to improved use and visualization of forest plots.

14: Robust Outlier Detection with Skewness-Adjusted Fences: Theoretical Foundations and Applications

Yunchae Jung, Minsu Park

Department of Statistics and Data Science, Chungnam National University, Republic of Korea

Outlier detection plays a crucial role in statistical analysis by ensuring data integrity and improving the reliability of inferences. Traditional methods, such as Tukey's boxplot, often struggle with skewed distributions, leading to inaccurate detection and potential misinterpretation of results. While approaches like the adjusted boxplot (Hubert and Vandervieren, 2008) provide some improvements, they can be computationally demanding and less effective under extreme skewness.

In this study, we present an outlier detection framework that incorporates a skewness-adjusted fence into an enhanced boxplot design. By utilizing a robust skewness measure based on the median absolute deviation, this method addresses key limitations of existing approaches, offering a computationally efficient and statistically reliable alternative for skewed distributions. Simulation studies and real-world applications demonstrate that the proposed method consistently improves detection accuracy while maintaining efficiency.

Additionally, we extend this approach to time-dependent data, showing its effectiveness in identifying outliers in time series settings. This extension makes the method applicable to a wide range of fields, including finance, healthcare, and environmental monitoring, where detecting anomalies in structured and evolving datasets is essential.

Keywords: Robust outlier detection, Skewness-adjusted boxplot, Influence function, Median absolute deviation

15: Minimum Area Confidence Set Optimality for Simultaneous Confidence Bands for Percentiles in Linear Regression: An Application to Estimating Shelf Life

Lingjiao Wang¹, Yang Han¹, Wei Liu², Frank Bretz^{3,4}

¹Department of Mathematics, University of Manchester, UK

²School of Mathematical Sciences and Southampton Statistical Sciences Research Institute, University of Southampton, UK

³Novartis Pharma AG, Basel, Switzerland

⁴Section for Medical Statistics, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Austria

Background The stability of a drug product over time is a critical property in pharmaceutical development. A key objective in drug stability studies is to estimate the shelf-life of a drug, involving a suitable definition of the true shelf-life and the construction of an appropriate estimate of the true shelf-life. Simultaneous confidence bands (SCBs) for percentiles in linear

regression are valuable tools for determining drug shelf-life in drug stability studies.

Methods In this paper, we propose a novel criterion, the Minimum Area Confidence Set (MACS), for identifying the optimal SCB for percentile regression lines. This criterion focuses on the area of the constrained regions for the newly proposed pivotal quantities, which are generated from the confidence set for the unknown parameters of a SCB. We employ the new pivotal quantities to construct exact SCBs over any finite covariate intervals and use the MACS criterion to compare several SCBs of different forms. Additionally, we introduce a computationally efficient method for calculating the critical constants of exact SCBs for percentile regression lines.

Results The optimal SCB under the MACS criterion is demonstrated to effectively construct interval estimates of the true shelf-life. The proposed method for calculating critical constants significantly improves computational efficiency. A real-world drug stability dataset is used to illustrate the application and advantages of the proposed approach.

16: One-Sided Simultaneous Tolerance Intervals Based on Kernel Density Estimates

Gian Louise Roy

University of the Philippines Diliman

Tolerance intervals are informative tools with wide-ranging applications in various fields, especially in laboratory medicine. They are valuable in medical decision making as they contain a specified proportion of values of the sampled population with high degree of confidence. When several biochemical analytes are measured from patients, simultaneous inference becomes useful. This study proposes nonparametric methods that construct simultaneous tolerance intervals (STIs) under the one-sided case. As most medical data show skewness and come from unknown underlying distributions, the proposed STIs are based on kernel density estimates. The methodologies used are evaluated by examining performance metrics, such as estimated coverage probabilities and expected lengths, and by comparing them with the usual Bonferroni-correction approach (BCA). The proposed methods show accurate results as the said metrics exhibit desirable patterns, with a few exceptions that are further examined and justified. These methods also address a spurious behavior that BCA results tend to display. The proposed one-sided nonparametric STIs are generally favourable than the ones

from BCA and can be improved through recommended future work that are laid out.

17: Robust Large-Scale Multiple Testing for Hidden Markov Random Field Model

Donghwan Lee¹, Jiyn Sun²

¹Department of Statistics, Ewha Womans University, Republic of Korea

²Integrated Biostatistics Branch, Division of Cancer Data Science, National Cancer Center, Republic of Korea

The hidden Markov random field model (HMRF), as an effective model to describe the local dependence of two or three-dimensional image data, has been successfully applied to large-scale multiple testing of correlated data, image segmentation, graph discovery, and so on. Given the unobservable random field, the emission probability (conditional distribution of observations) is usually assumed to be known, and the Gaussian distribution is frequently used. To achieve robustness, we introduce a novel framework for large-scale multiple testing when the emission probability distribution of HMRF is unknown or misspecified. We build the inferential procedure for estimating parameters and the false discovery rate (FDR) based on a quadratically convergent method for computing non-parametric maximum likelihood estimates of a mixing distribution. Furthermore, we integrate latent variable modeling with the knockoff filter method to improve FDR control in testing. The proposed method is validated by simulation studies, which show that it outperforms the other existing methods in terms of FDR validity and power. A real data example for neuroimaging is illustrated to demonstrate the utility of the proposed procedure.

18: Model Informed Assurance Approach for 3-Way PK Similarity Studies

Rachid El Galta, Roland Baumgartner

Sandoz, Germany

In the absence of actual data, published pharmacokinetic (PK) models can simulate subjects' PK profiles to estimate geometric mean ratios and coefficient of variations for parameters like AUC and Cmax. These estimates can be used to inform sample size calculations for PK similarity studies. However, the accuracy depends on the quality of the PK model and input parameters. Ignoring uncertainty can lead to underpowered studies.

To address this, we use an assurance approach alongside power calculations. This involves simulating PK profiles with a published PK model, considering parameter uncertainty by sampling from a multivariate normal distribution. We generate multiple parameter sets, simulate PK profiles by treatment arm for each, and perform equivalence testing. Assurance is the proportion of successful equivalence tests.

Combining assurance with traditional power calculations provides a more comprehensive assessment of sample size considerations.

19: Korea Sequence Read Archive (KRA) - A Public Repository for Archiving Raw Sequence Data

Jaeho Lee

KRIBB, Korea, Republic of (South Korea)

The Korea Sequence Read Archive (KRA; <https://kbds.re.kr/KRA>) is a publicly available repository of high throughput sequencing data as a part of the Korea BioData Station (KBDS; <https://kbds.re.kr/>) database. KRA collects and provides key nucleotide sequence data, including files in FASTQ or FASTA format and rich metadata generated by various NGS technologies. The primary objective of the KRA is to support and promote the use of nucleotide sequencing as an experimental research platform. It achieves this by offering comprehensive services for data submission, archiving, searching, and downloading. Recently, the existing collaboration with DDBJ has been further strengthened to establish close cooperation with INSDC. As a result, KRA now supports data submission to INSDC via DDBJ DRA, and through enhanced browser functionalities, users can search and download data more efficiently. By ensuring the long-term preservation and accessibility of nucleotide sequence data and through continuous development and improvements, KRA remains an important resource for researchers utilizing nucleotide sequence analysis data. KRA is available at <https://kbds.re.kr/KRA>.

20: Integrative Analysis of Transcriptomic and Epigenomic Dynamics of Liver Organoids using Single Cell RNA-Seq and ATAC-Seq

Kwang Hoon Cho, Jong-Hwan Kim, Jimin Kim, Jahyun Yun, Dayeon Kang

Korea Research Institute of Bioscience and Biotechnology, Korea, Republic of (South Korea)

Previously, we developed a novel method to generate functionally mature human hepatic organoids derived from pluripotent stem cells (PSCs), and their maturation was validated through bulk RNA sequencing (RNA-seq). In this study, we aimed to characterize the heterogeneity and dynamic changes in the transcriptome and epigenome at the single-cell level. To achieve this, we employed single-cell RNA sequencing (scRNA-seq) and single-cell ATAC sequencing (scATAC-seq) using the 10x Chromium platform.

Hepatic organoids were cultured under two distinct medium conditions: hepatic medium (HM) and differentiation medium (DM). A total of 39,310 and 36,940 individual cells were analyzed using scRNA-seq and scATAC-seq, respectively. To validate our findings, we compared our data with publicly available RNA-seq datasets from liver organoids and liver tissues at various stages of differentiation, including induced pluripotent stem cells (iPSCs), DM-treated cells, primary human hepatocytes (PHHs), and adult liver tissues.

Our analysis revealed that cells clustered into 10 to 11 distinct subpopulations, representing different developmental stages in both scRNA-seq and scATAC-seq datasets. Furthermore, integrative analysis of scRNA-seq and scATAC-seq data identified coordinated changes in gene expression and chromatin accessibility near key liver differentiation marker genes. These findings indicate that hepatic organoids cultured under HM and DM conditions consist of heterogeneous cell populations spanning multiple stages of hepatic differentiation.

In conclusion, single-cell transcriptomic and epigenomic profiling provided insights into the cellular diversity and developmental trajectory within hepatic organoids. This study highlights the utility of scRNA-seq and scATAC-seq in elucidating the molecular dynamics underlying liver differentiation and maturation.

21: Leveraging Tumor Imaging Compositional Data Structure in Model Feature Space for Predicting Recurrence in Colorectal Carcinoma

Olivia J Bobek, Nicholas Larson, Rish K Pai, Fang-Shu Ou

Mayo Clinic, United States of America

Background/Introduction The quantitative segmentation algorithm QuantCRC extracts morphologic features of digitized H&E slides in colorectal carcinoma (CRC), quantitatively

decomposing the tumor bed area into stroma and stromal subtypes, necrosis, and tumor components. These features have previously been incorporated as linear predictors in a LASSO regularized regression model for cancer recurrence in a cancer registry study. However, as compositional data, representing these features as simple proportions may not maximize their informativeness for prediction. Likewise, algorithms based on linear predictors may fail to account for more complex relationships between compositional features and outcome. The objective of this research was to investigate how commonly used log-ratio transformations for compositional data impact QuantCRC-based prognostic modeling performance as well as assess competing machine learning algorithms that may offer benefits for compositional feature spaces.

Methods The study cohort consisted of 2411 CRC patients from the Colon Cancer Family Registry. The outcome of interest was recurrence-free survival, measured as time from surgery to recurrence or last follow-up. The original LASSO model included 15 QuantCRC features, tumor stage (I-IV) and mismatch repair status (deficient vs. proficient). The proposed model feature space included the additive log-ratio transformations of the composition variables in addition to the clinical variables, yielding 34 features total. In addition to LASSO, elastic net and gradient boosting machine (GBM) algorithms were also applied using the log-ratio feature set. Training was performed using 10-fold cross validation on 80% ($n=1928$) and tested on 20% ($n=483$) of the data. Harrell's C-index was used to assess discrimination.

Results On the training set, the original LASSO produced a Harrell's C-index of 0.697 (bootstrapped 95% Confidence Interval (CI): 0.672, 0.723) and the LASSO with log-ratio features produced a C-index of 0.703 (95% CI: 0.679, 0.729). The C-index for the elastic net and GBM was 0.704 (95% CI: 0.677, 0.731) and 0.719 (95% CI: 0.692, 0.744) respectively. In the test data, the LASSO with the log-ratio transformation produced a slightly improved C-index: 0.701 (95% CI: 0.650, 0.746) compared to the original features (0.697 (95% CI: 0.646, 0.743)). The elastic net resulted in a C-index of 0.703 (95% CI: 0.653, 0.749) and GBM produced a C-index of 0.702 (95% CI: 0.647, 0.751).

Conclusion The additive log-ratio transformation is a compositional data representation to consider for predictive models. In this application, feature engineering based on compositional structure slightly improved model performance. All algorithms with compositional data features demonstrated comparable model discrimination.

22: BayesPIM: A Bayesian Prevalence-Incidence Mixture Model for Screening Outcomes, with an Application to Colorectal Cancer

Thomas Klausch, Birgit Lissenberg-Witte, Veerle Coupé

Amsterdam University Medical Center

Background

Screening programs for diseases, such as colorectal cancer (CRC), involve inviting individuals in regular or irregular intervals for a test, such as the Fecal immunochemical test (FIT) or a colonoscopy. The resulting data can be analyzed to obtain the time to (pre-state) disease which, when additionally regressed on covariates, such as age and gender, is informative on risk heterogeneity. Such information helps decide whether screening intervals should be personalized to identified risk factors.

We present the R package BayesPIM – Bayesian prevalence-incidence mixture model – which is particularly suited in settings where individuals are periodically tested (interval censoring), have the disease at baseline (prevalence), baseline tests may be missing, and the screening test has imperfect sensitivity. We motivate the model using data from high-risk familial CRC surveillance through colonoscopy, where adenomas, precursors of CRC, are the primary target of screening. Besides demonstrating the functionalities of BayesPIM, we also show how to evaluate model performance using simulations based on the real-world CRC data.

Methods BayesPIM models the interval-censored time to incidence via an accelerated failure time model while handling latent prevalence, imperfect test sensitivity, and covariate data. Internally, a Metropolis-within-Gibbs sampler and data augmentation is used, implemented through an Rcpp backend. A user-friendly R interface is available. Model fit can be assessed using information criteria and validated against a non-parametric estimator of cumulative incidence.

Additionally, performance is evaluated by resampling the real-world CRC screening data. Specifically, we set the data-generating model parameters to their estimates and then generate screening times and outcomes that closely resemble those observed in practice via an innovative algorithm. Repeatedly comparing estimates on these resampled datasets to the true values assesses model performance under realistic data conditions.

Results In the CRC application, baseline prevalence of adenomas was estimated at 27.4% [95% CI: 22.2%, 33.3%], with higher prevalence in males and older individuals. Among those free of adenoma at baseline, incidence reached 20% at five years and 45% at ten years, with older individuals experiencing faster incidence. Resampling simulations based on the CRC data showed that model estimation remained stable if informative priors on test sensitivity were imposed, even at low sensitivity (40%).

Conclusion BayesPIM offers robust estimation of both prevalence and incidence under complex, real-world screening conditions, including uncertain test sensitivity, latent disease status, and irregular intervals. The model demonstrated stable performance under varying test sensi-

tivities, highlighting its practical value for designing more effective, patient-centered screening programs.

23: Joint Modelling of Random Heterogeneity in Longitudinal and Multiple Time-to-Events in Colon Cancer

Divya Dennis, Jagathnath Krishna Km

Regional Cancer Centre, Thiruvananthapuram, Kerala, India, India

Background In cancer survival studies, disease progression can be assessed with longitudinal study designs where the patients are observed over time and the covariates information (biomarkers, carcinoembryonic antigen -CEA) are measured repeatedly during the follow up period. Apart from repeated measured covariates, multiple survival outcomes were observed longitudinally. Also there may exist unobserved random heterogeneity between the survival outcomes. This motivated to derive a joint multi-state frailty model (JMF) capable of predicting the risk for multiple time-to events simultaneously utilizing the dynamic predictors and random heterogeneity factor, the frailty. The frailty variable was assumed to follow gamma distribution, thus forms the joint multi-state gamma frailty model (JMGFM).

Methodology: For accounting heterogeneity, longitudinal outcome and multiple time-to-events, we derived a JMGFM. The longitudinal sub-model was modeled using linear mixed model and the survival sub-model using multi-state gamma frailty model (MGFM). The latent variable was used to link the longitudinal and multiple time-to-event sub-models. The parameters were estimated using maximum likelihood estimation method. The existing MGFM and the developed model were illustrated using colon cancer patient data. The covariate considered for risk prediction were, composite stage, lymph node involvement, T4, age, sex, PNI, LVE; and CEA as longitudinal outcome.

Results The study observed that frailty coefficient had a significant impact on predicting the risk at each transition states along with longitudinally measured covariate. So JMGFM were found to be better predictive than MGFM. The JMGFM model is capable of providing dynamic risk prediction simultaneously, which the MGFM cannot. The present study identified that PNI (transition from diagnosed as disease to death), Composite Stage (transition from recurrence to death; transition from metastasis to death and transition from recurrence to metastasis) lymph-node involvement and age along with the longitudinally measured CEA value as significant prognostic factors for predicting the multiple time-to-events based on the proposed JMGFM and also found that for each transition state, the longitudinal observation (CEA) has strong association with corresponding survival events (η ranges from 1.3 to

1.5).

Conclusion Thus we conclude that the joint multi-state frailty model as a better model for simultaneous dynamic risk prediction of multiple events in presence of random heterogeneity in longitudinal study design.

Keywords: *Multi-state model, Joint multi-state model, joint multi-state frailty model, longitudinal sub-model, Colon cancer*

24: Refining the Association Between BMI, Waist Circumference, and Breast Cancer Risk in Postmenopausal Women using G-Formula Method

Somin Jeon^{1,2}, Boyoung Park^{1,3}, Junghyun Yoon^{1,2}

¹Department of Preventive Medicine, Hanyang University College of Medicine, Seoul, Republic of Korea

²Institute for Health and Society, Hanyang University, Seoul, South Korea

³Hanyang Institute of Bioscience and Biotechnology, Hanyang University, Seoul, Republic of Korea

Purpose. Previous studies have shown an increased risk of postmenopausal breast cancer (BC) in obese women. However, these studies did not focus on longitudinal changes in obesity levels and did not account for time-varying covariates. This study applies the g-formula method to assess how changes in BMI and waist circumference associate with subsequent BC risk.

Methods Data were obtained from the Korean National Health Insurance Database. We utilized data from the national BC screening, with baseline data including women who underwent screening in 2009-2010. Screening information in the subsequent biennial cycles (2011-2012 until 2019-2020) was examined, and only women with postmenopausal status at baseline and with at least three screenings were included in the analysis. Incident BC cases were ascertained until 2021. We applied the g-formula method to compare BC risk in women who maintained a certain BMI/waist circumference level versus the natural course. Hazard ratios (HRs) were estimated, and the model was adjusted for age, fixed covariates, and time-varying covariates.

Results Of the 91,092 postmenopausal women, the mean (SD) age was 60.7 (7.5), and the mean (SD) BMI and waist circumference were 24.3 (3.1) and 79.9 (8.1), respectively. Results from the G-formula show that compared to women with a natural course of BMI, those who

maintained a normal BMI level (<23) or overweight BMI (23 to <25) had a decreased BC risk (adjusted hazard ratio [aHR] 0.93, 95% CI 0.90 – 0.95, and aHR 0.97, 95% CI 0.96–0.98, respectively). In contrast, those who maintained obese status had an increased BC risk (obese 1, BMI 25 to <27.5 , with an aHR of 1.07, and obese 2, BMI 27.5, with an aHR of 1.20). A similar pattern was observed in the results for waist circumference.

Conclusions Results from the g-formula indicate that maintaining a normal BMI or waist circumference is associated with a lower BC risk, while obese women are at an increased risk of postmenopausal breast cancer.

Acknowledgments: This study was funded by the National Research Foundation of Korea (NRF) (grant no. RS-2023-00241942, RS-2024-00462658, and 2021R1A2C1011958).

25: Building Cancer Risk Prediction Models by Synthesizing National Registry and Prevention Trial Data

Oksana Chernova¹, Donna Ankerst^{1,2}, Ruth M Pfeiffer³

¹Technical University of Munich, Germany

²Department of Urology, University of Texas Health Science Center at San Antonio, USA

³Biostatistics Branch, National Cancer Institute, NIH, HHS, Bethesda, USA

Current online United States (US) five-year prostate cancer risk calculators are based on screening trials or databases not calibrated to the heterogeneous US population. They are underpowered for the rarer outcome of high-grade disease, particularly for the subpopulation of African Americans, who are underrepresented in many national trials. The US National Cancer Institute (NCI) Surveillance, Epidemiology, and End Results (SEER) program has monitored state cancer rates since 1973, more recently adding Gleason grade. SEER rates are stratified by five-year age groups and race, filling in statistical power gaps for African Americans. This talk provides the statistical method for integrating SEER incidence and mortality rates with time-to-event data with competing risks from prevention and screening trials following the NCI Breast Cancer Risk Assessment Tool. The methodology allows development of a contemporary 5-year high-grade prostate cancer risk prediction model that is trained from merging individual-participant data from the Selenium and Vitamin E Cancer Prevention Trial (SELECT) with population aggregated data in SEER. Simulation of a contemporary US validation set is performed by merging individual-level data from the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (PLCO) with SEER.

26: Modelling Individual-Level Uncertainty from Missing Data in Personalised Breast Cancer Risk Prediction

Bethan L. White, Lorenzo Ficarella, Xin Yang, Douglas F. Easton, Antonis C. Antoniou

University of Cambridge, United Kingdom

Breast cancer risk prediction models use a range of predictors to estimate an individual's chance of developing breast cancer in a given timeframe. These can facilitate risk stratification, to identify individuals who would benefit most from screening or preventive options. The BOADICEA breast cancer risk model, implemented in the CanRisk tool (www.canrisk.org), uses genetic, lifestyle, hormonal, family history and anthropometric data to estimate an individual's risk. When implementing risk prediction models, risk predictor data are often incomplete. Point-estimates calculated when some risk factor data are missing can therefore hide considerable uncertainty.

We developed a methodological approach for quantifying uncertainty and the probability of risk-reclassification in the presence of missing data. We employed Monte Carlo simulation methods to estimate the distribution of breast cancer risk for individuals with missing data, using multiple imputation by chained equations (MICE) with UK Biobank and KARMA as reference datasets to sample missing covariates. We developed a framework for estimating uncertainty, that can be applied to any given individual with missing risk factor data. We used exemplar cases to assess the probability that collecting all missing data would result in a change in risk categorisation, on the basis of the 10-year predicted risk from age 40, using the UK National Institute for Health and Care Excellence (NICE) guidelines.

For example, a woman whose mother and sister have both been previously diagnosed with breast cancer, but with all other personal risk factor information unmeasured, will be categorised as at "moderate risk" by the BOADICEA model, with around a 5% chance of developing breast cancer between the ages of 40 and 50. However, if all remaining risk factor information were measured, our methodology estimates a 52% chance of reclassification to the "population risk" group, and a 5% chance of reclassification to the "high risk" group. Taking into account all missing risk factor information, an estimated 95% uncertainty interval for the risk point estimate would be (0.9%, 9.0%).

These results demonstrate that there can be a considerable likelihood of reclassification into a different risk category after collecting missing data. The methodology presented in this work can identify situations where it would be most beneficial to collect additional patient information before making decisions in clinical settings.

27: Time-Varying Covariates in Survival Analysis: a Graphical Approach to Assessing the Risk of Cardiovascular Events and Aortic Valve Sclerosis Development

Arianna Galotta¹, Francesco Maria Mattio¹, Veronika Myasoedova¹, Elisabetta Salvioni¹, Paolo Poggio^{1,3}, Piergiuseppe Agostoni^{1,2}, Alice Bonomi¹

¹Centro Cardiologico Monzino, IRCCS, Milan, Italy

²Department of Clinical and Community Sciences, University of Milan, Milan, Italy

³Department of Biomedical, Surgical and Dental Sciences, University of Milan, Milan, Italy

Background Survival analysis is essential for studying the time to the occurrence of an event of interest, such as death or the onset of a disease. When covariates change over time, it is crucial to consider these variations to estimate the relationship between exposure and outcome accurately and robustly. While using the Cox model with time-dependent covariates is methodologically appropriate, its graphical representation remains challenging. This study focuses on evaluating the development of aortic valve sclerosis (AVSc) as an exposure condition to the risk of cardiovascular (CV) events, taking into account its progression over time.

Methods The relative risk of CV events linked to AVSc development was assessed using the Cox proportional hazards model. To generate the survival curves, we applied the method proposed by Simon and Makuch (Schultz et al., 2002). This approach differs from the traditional Kaplan-Meier method, which treats covariates as fixed; the difference imposed by considering a time-dependent covariate is in the interpretation of the risk set. In our case, a time-varying covariate leads to a continuous renewal of risk sets based on the value of the covariate at each time point. Therefore, the risk set includes all individuals at risk just before time t , whose covariate value indicates their membership in the relevant group at that time.

Results Time-dependent analyses were conducted to evaluate AVSc development as a time-sensitive exposure to CV events. Participants who developed AVSc during the follow-up period were considered unexposed from baseline until the onset of development, after which they were classified as exposed. The hazard ratio related to the AVSc development was then evaluated using the Cox proportional-hazards model. The analysis with the time-dependent covariate approach provided a more detailed understanding of the association between the AVSc development and the risk of CV events over time. The survival curves generated using this method demonstrated that accounting for the time-varying nature of AVSc exposure significantly impacted the prognosis of patients.

Conclusion This study emphasises the importance of considering time-varying covariates in

survival analysis for an accurate risk estimate. Although the Cox model with time-dependent covariates is the correct methodological choice, its graphical representation is complex. The method proposed by Simon and Makuch enhances the traditional Kaplan-Meier approach by allowing the integration of covariates that evolve over time. This is particularly relevant in medical research, where dynamic exposures must be considered to avoid misleading conclusions.

28: Polygenic Scores as Tools for Intervention Selection in the Setting of Finasteride for Prostate Cancer Prevention

Allison Meisner

Fred Hutchinson Cancer Center, United States of America

Background/introduction: Polygenic risk scores (PRS) have been proposed as tools for intervention selection. PRS are weighted combinations of single nucleotide polymorphisms (SNPs) where each SNP is weighted by its association with outcome risk. An alternative approach utilizes predictive polygenic scores (PPS), in which the weight for each SNP corresponds to its association with intervention effect. We compare the utility of PRS and PPS for identifying individuals expected to benefit from finasteride in the prevention of prostate cancer.

Methods: We used data from the Prostate Cancer Prevention Trial (PCPT), a randomized trial of finasteride for prostate cancer prevention. Of the 8,506 men with available genotype data, 1,440 developed prostate cancer. We used the Polygenic Score Catalog to identify a recently developed prostate cancer PRS. We split the data into training (2/3 of the data) and test (1/3 of the data) sets. We constructed three scores, each of which was a combination of 198 SNPs in the PRS published on the Polygenic Score Catalog: (1) a PRS based on the coefficients published in the Polygenic Score Catalog (PRS1), (2) a PRS based on coefficients estimated in the training data via logistic regression (PRS2), and (3) a PPS based on the interaction between each SNP and randomization to finasteride, estimated in the training data via logistic regression. In the test data, we compared the three scores based on the reduction in the rate of prostate cancer when a given score is used for intervention selection.

Results: In the test data, 17.0% of men developed prostate cancer and finasteride was significantly associated with a reduction in risk of prostate cancer; thus, the default setting is to treat all men with finasteride. For PRS1, there was no threshold at which treatment with finasteride would not be recommended; thus, use of PRS1 to guide intervention use would not reduce the rate of prostate cancer. For PRS2, 0.2% of men would not be recommended

finasteride, leading to a reduction in the rate of prostate cancer of < 0.001%. Finally, for the PPS, 35.3% of men would not be recommended finasteride, leading to a reduction in the rate of prostate cancer of 3.0% if the PPS were used to guide intervention use.

Conclusion: In this analysis of PCPT data, PPS demonstrated substantially greater clinical utility as tools for intervention selection compared to PRS. PPS should be considered as tools for intervention selection more broadly.

29: Implementation of a Disease Progression Model Accounting for Covariates

Gabrielle Casimiro, Sofia Kaisaridi, Sophie Tezenas du Montcel

ARAMIS, Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, CNRS, Inria, Inserm, AP-HP, Groupe Hospitalier Sorbonne Université, Paris, France

Introduction Disease progression models are promising tools for analysing longitudinal data presenting multiple modalities. Such models can be used to estimate long-term disease progression and reconstruct individual trajectories. Inter-patient variability is often modeled as random perturbations around a fixed reference. However, much of this variability is driven by external factors such as genetic mutations, gender, level of education or socio-economic status.

In this work, we extend a non-linear mixed-effects disease progression model (Disease Course Mapping Model), implementing a multivariate logistic framework to explicitly account for covariates. We illustrate the potential of this approach by modelling the evolution of CADASIL disease, the most frequent small artery brain disease caused by pathogenic variants of the NOTCH3 gene, using the genetic mutation location as a covariate.

Methods A general formulation involves a non-linear mapping η between timepoints and clinical markers, parametrized by fixed effects α (population level) and random effects β_i (individual level): $y_i = \eta(\alpha | \beta_i)$.

The Disease Course Mapping Model applies time reparameterization to realign all individual trajectories into a common timeline, accounting for spatiotemporal variability. To do so, it estimates a population parameter expressing the average disease onset time, enabling direct comparison of features (such as scores or biomarker values measured longitudinally) at this time and identifying the sequence of symptom onset.

To incorporate baseline covariates in the model, the existing paradigm was extended. Instead

of estimating a fixed effect α parametrizing the average disease course, we introduced a link function f_ϕ that can predict an expected trajectory of the disease conditioned by a given set of covariates c_i .

Results The proposed model has been implemented in the Open-source library Leaspy. Applied to different clinical scores, it reveals significant differences according to the mutation location: patients with the pathogenic variant located in EGFr domains 1-6, previously identified as a determinant of disease severity, showed a faster and more pronounced decline in the Rankin score assessing the severity of disability.

Conclusion This approach allows us to explicitly model how external factors influence disease progression rather than treating variability as purely stochastic. While the current model incorporates a single binary covariate, future work will focus on extending this framework to handle multiple covariates simultaneously and to integrate continuous variables.

30: Genetics Influences LDL-C Response to Statin Therapy: Short- and Long-Term Observational Study with Functional Data Analysis

Andrea Corbetta^{1,2,3}, Emanuele Di Angelantonio^{1,4}, Andrea Ganna³, Francesca Ieva^{1,2}

¹Human Technopole, Milan, Italy

²Politecnico Di Milano, Milan, Italy

³Institute for Molecular Medicine Finland, Helsinki, Finland

⁴University of Cambridge, Cambridge UK, UK

Introduction Understanding the genetic basis of lipid-lowering responses to statin therapy may provide critical insights into personalized cardiovascular treatment strategies. This study employs advanced statistical methods to investigate how genetic predisposition, captured through polygenic scores (PGS) for low-density lipoprotein cholesterol (LDL-C), influences short-term and long-term changes in LDL-C levels following statin initiation.

Methods We utilized data from the FinnGen cohort, focusing on LDL-C measurements in two distinct groups: (1) a short-term group of 11,343 individuals with LDL-C measurements recorded within one year before and after initiating statin therapy and (2) a long-term group of 15,864 individuals who had maintained statin therapy for a minimum of five years. The LDL-C trajectories were modelled as functional objects, allowing us to apply functional principal components analysis (FPCA) to identify independent patterns of LDL-C response.

In the short-term group, we modelled the absolute and relative reduction of LDL-C using linear regression models with PGS as a predictor. In the long-term group, we analyzed the first two FPCA components: the first principal component (PC1) representing the baseline LDL-C level (mean pattern) and the second principal component (PC2) capturing the LDL-C reduction pattern. Genome-wide association studies (GWAS) were conducted to identify genetic variants associated with these phenotypic patterns, applying stringent Bonferroni correction for multiple testing.

Results We observed that individuals in the highest PGS tertile experienced a greater absolute LDL-C reduction in the first year after statin initiation, with a mean reduction of 8.12 mg/dL (95% CI: 6.93–9.57) compared to the lowest tertile. However, this group demonstrated a smaller relative reduction of 1.81% (95% CI: 0.06–2.99). In the long-term group, higher PGS was associated with elevated LDL-C levels over five years but no significant association was found between PGS and LDL-C change patterns. The GWAS identified significant genome-wide loci for relative LDL-C reduction and baseline LDL-C levels, with lead variants near genes previously implicated in lipid metabolism.

Conclusion Our findings suggest that short-term LDL-C response exhibits a genetic basis strongly linked to baseline LDL-C regulation. In contrast, long-term LDL-C changes appear predominantly influenced by non-genetic factors such as adherence. Nonetheless, individuals with higher LDL-C PGS consistently maintain higher LDL-C levels over extended periods. These results underscore the complex genetic architecture of LDL-C response to statins and highlight the utility of FPCA in characterizing dynamic lipid trajectories.

31: Longitudinal Analysis of Imprecise Disease Status using Latent Markov Models: Application to Italian Thyroid Cancer Observatory Data

Silvia D'Elia¹, Marco Alfò¹, Maria Francesca Marino²

¹Sapienza University of Rome (Italy)

²University of Florence (Italy)

Background Longitudinal data are widely used in medicine to monitor patients over time, providing dynamic view of disease progression, and treatment responses. Ordinal scales are often used to measure response to treatment or summarise disease severity.

Methods Latent Markov (LM) models represent an important class for dealing with ordinal longitudinal data. LM models are based on a latent process which is assumed to follow a Markov chain with a certain number of states (latent states). The state characterises the

(conditional) response distribution at each time occasion.

LM model allows to analyse longitudinal data when dealing with:

- measurement error
- unobserved heterogeneity

Such models estimate transition probabilities between latent states, also including individual covariates¹. Of particular interest is the evaluation of patient dynamics over time as a function of individual covariates (both constant and time-dependent).

Application

A latent Markov model is used to analyse data from the Italian Thyroid Cancer Observatory (ITCO), a database of over 15,000 patients with diagnosis of thyroid cancer, treated in different clinical centres in Italy. Despite the high survival rate, the risk of recurrence remains significant, and long-term monitoring is needed to detect recurrence early and maintain appropriate therapies². The study aims to monitor and assess the effectiveness of the response to treatment over the time, trying to identify factors that predict true disease status.

Patients are monitored prospectively from the date of surgery, with follow-up visits at 12, 36, and 60 months. At each follow-up visit, the response to treatment is assessed through clinical, biochemical, and imaging findings and response is classified into 4 categories: excellent (ER, no evidence of disease), indeterminate (IND), biochemical incomplete (BIR) and structural incomplete (SIR, evidence of disease). However, this classification has limitations: categories synthesise multiple measurements prone to error, are influenced by unobserved factors, and the disease status (evidence vs. no evidence of disease) is not directly observable due to the presence of ambiguous categories (IND, BIR).

While around 50% of patients show clearly no evidence of disease at any time point, 30–40% fall into indeterminacy area.

Conclusion Latent Markov models may lead to a better understanding of patients' clinical trajectories depicting a more accurate picture of patients' dynamics, considering variables that may influence transitions between states.

¹Bartolucci et al., **Latent Markov Models for Longitudinal Data, 2012.**

²Haugen et al., **2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer, 2016.**

32: Long-Term Risk Prediction from Short-Term Data – a Microsimulation Approach

Moritz Pamminger, Theresa Ullmann, Moritz Madern, Daniela Dunkler, Georg Heinze

Medical University of Vienna, Austria

Background

In medical research, long-term risk prediction is often desirable, e.g. to predict the risk of a cardiovascular or other health event within the next 30 years. To estimate such a prediction model requires data with a long enough follow-up. Such data are rarely available and may be outdated. Our aim is to develop and evaluate methods to harness contemporary data for long-term predictions.

Methods

We assume longitudinal data with 2-5 possibly irregular measurements of 5-20 prognostic factors (e.g. cholesterol, blood pressure etc.) per individual over a 5-years period and associated survival outcomes. We present a microsimulation-based strategy to obtain predictions of survival and of trajectories of the prognostic factors over a long-term prediction horizon of 20-30 years.

First, we trained models using the current values of prognostic factors to predict subsequent measurements and the event status with a short-term prediction horizon of 1-2 years. Starting with individual-specific initial values of the prognostic factors, these short-term models were then applied to generate follow-up values of the prognostic factors and of the survival state as draws from the respective predictive distributions. These values serve as new baseline for the next prediction-and-generation step. Iteration proceeds until an event is predicted or the long-term prediction horizon is reached. For each individual multiple (e.g. 1.000) trajectories for prognostic factors and the survival process are generated, which can be suitably summarized.

We validated the approach using various synthetic datasets for which long-term follow-up was available. We artificially censored these datasets to mimic data with short-term follow-up, which we used to train our models. Then we applied the microsimulation approach to make long-term predictions and compared the predicted outcomes with the observed ones in the training set. We also validated predictions in an independent test set.

Results

The approach was implemented in an R package for convenient application in various situations. The package provides flexible options to specify short-term models. It can perform predictions for individuals, efficiently processing entire datasets, and present results with appropriate graphical summaries.

Conclusion

Despite some limitations, the method effectively handles irregular time intervals in the training data and allows capturing nonlinear and interaction effects for prognostic factors and survival. It provides analysts with a flexible tool for long-term prognosis across various fields and in the future may provide a practically useful framework for individual long-term prognosis at routine health screenings. This work was supported through the FWF project P-36727-N.

33: Deep Learning Algorithm for Dynamic Survival Prediction with Competitive Risks

Tristan Margat  ^{1,2,3}, Marine Zulian¹, Agathe Guilloux², Sandrine Katsahian^{2,3,4}

¹Healthcare and Life Sciences Research, Dassault Systemes, France

²HeKa team, INRIA, Paris, France

³Universit   Paris Cit  , France

⁴URC HEGP, APHP Paris

Background The medical follow-up of a patient suffering from cancer is spread over time, making it possible to obtain repeated measurements that allow to consider the progression of the disease state over time. The development of a prognostic solution requires the ability to update predictions of the occurrence of clinical events over time according to new measurements, i.e., to make dynamic predictions.

In oncology, patients can face appearance of metastasis, other diseases due to comorbidities or death. It can be useful to predict which of these events will occur first; this is referred in the literature as competing risks. We aim to develop new methodologies capable of considering longitudinal data for predicting competing risks.

Methods:

Various Deep Learning algorithms have recently been developed to consider longitudinal survival data and competing risks [1] [2]. However, as they consider the entirety of the patient's available longitudinal data to create a time-independent static embedding, they suffer from bias when used to predict survival over the patient's follow-up interval. Recent methodologies use a progressive approach to integrate patient's data [3], allowing to get an

embedding of features that varies over time. It shows superior results in a context of classical survival analysis.

We have chosen to use this type of methodology, modifying the method used to create the embedding of longitudinal data, and to extend it to competing survival setting.

In addition, we have developed a new simulation scheme to obtain synthetic longitudinal survival data with competitive risks.

Results/Conclusion:

We will present results on different approaches to consider both longitudinal and survival data and their limitations in order to produce unbiased predictions. We compare our algorithm with existing algorithms on simulated data and a subset of real-world data from the Framingham Heart Study whose aim was to study the etiology of cardiovascular disease.

References

- [1] Lee, C. et al. (2019). Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Transactions on Biomedical Engineering*
- [2] Moon, I. et al. (2022). SurvLatent ODE: A Neural ODE based time-to-event model with competing risks for longitudinal data improves cancer-associated Venous Thromboembolism (VTE) prediction. In *Machine Learning for Healthcare Conference*
- [3] Bleistein, L et al. (2024). Dynamical Survival Analysis with Controlled Latent States. *arXiv preprint arXiv:2401.17077*.

34: Identifying Cutoff in Predictors in Survival Analysis: An Ensemble Strategy for Flexible Knot Selection

Stefania Lando, Giulia Lorenzoni, Dario Gregori

Unit of Biostatistics, Epidemiology and Public Health, Department of Cardiac, Thoracic, Vascular Sciences and Public Health, University of Padova, Padova, Italy

Background Restricted cubic splines (RCS) are widely used in Cox proportional hazards models to capture nonlinear relationship between a continuous biomarker and patient outcomes. Traditional approaches for knot selection often rely on predefined quantiles (e.g., 5th, 35th, 65th, and 95th percentiles), normally arbitrarily chosen, or a fixed number of knots systematically placed across the biomarker range. All such strategies present certain limitations: quantile-based methods provide stability and reproducibility but may oversimplify underlying nonlinear relationships, whereas fixed knots are at risk of overlooking variations

in the data.

Methods Our study explores an ensemble methodology that seeks to put together the robustness of quantile-based knot placement with the flexibility of data-driven strategies, aiming to provide robust knot placement while preserving clinical meaning in cutoff determination. The core idea is to maintain the intuitive simplicity of quantile-based knots while introducing a selective tuning mechanism—guided by cross-validation—to refine their placement. In parallel, our approach incorporates time-dependent ROC analysis to identify clinically relevant cutoffs for risk stratification at a chosen time horizon.

Applications

The proposed methodology can be applied in various clinical and epidemiological settings where risk stratification based on continuous biomarkers is essential. Examples include oncology for identifying prognostic thresholds in tumor markers, cardiology for refining cardiovascular risk scores, and infectious disease modeling for determining severity cutoffs. Additionally, this approach can be extended to precision medicine, where patient subgroups with distinct risk profiles can be identified for targeted interventions.

Conclusion By integrating flexible knot placement with an ensemble-based cutoff strategy, the method enhances the adaptability of spline-based Cox models while preserving clinical relevance.

35: Estimating Quality Adjusted Life Years (QALYs) from a Joint Modeling Framework: a Simulation-Based Study

Vincent Bonnemains¹, Yohann Foucher², Philippe Tessier¹, Etienne Dantan¹

¹1. Nantes Université, Univ Tours, CHU Nantes, INSERM, MethodS in Patients-centered outcomes and HEalth Research, SPHERE, F-44000 Nantes, France

²2. INSERM, CIC-1402, Centre Hospitalier Universitaire de Poitiers, Université de Poitiers, Poitiers, France.

Background. Clinical trials investigators often chose patient survival as the primary outcome.

When health-related quality of life (HRQoL) outcomes are considered, they are usually analyzed secondarily and separately from the survival outcome, precluding the consideration of potential trade-offs between them. In contrast, Quality-Adjusted Life Years (QALYs) are a

composite outcome that allows the two stakes to be considered simultaneously by weighting years of life by HRQoL indexes (utility scores) that reflect individual preferences. Hence, QALYs could be a practical primary outcome for assessing treatments benefit.

However, the estimation of QALYs usually relies on non-parametric approaches suffering several methodological pitfalls. This work aims to propose a sounder method for estimating QALYs using the shared random effects joint modelling framework.

Methods. We developed a shared random-effect joint model, the longitudinal utility scores being considered through a mixed beta regression and the time-to-death through a proportional hazard Weibull model. We then proposed a method for estimating QALYs using this model. We compared its performances with the commonly used non-parametric method through a simulation study.

We simulated a wide range of clinical trials considering the presence and absence of treatment effect, 200 and 500 included patients, one and two utility score measurements per patient per year for three years, and two censoring rates: 0% and 30% at a three-year horizon. We also considered different data generation mechanisms resulting in well-specified or misspecified models. For each scenario, we simulated 1000 data samples. The treatment effect was estimated in terms of QALYs at a three-year horizon.

Results. Our proposed method provided unbiased estimates of QALYs and significant improvements over the non-parametric approach when the joint model was well-specified. This was particularly the case when a low number of repeated utility measurements per patient or a high censoring rate was simulated. The two methods performed poorly when simulating the risk of event with non-proportional hazards.

Conclusions. We proposed a method based on joint modeling for estimating QALYs. We reported accurate estimations for clinical trials with moderate sizes when the model is well specified. However, we found the estimations to be sensitive to model misspecification. We are working to develop additional modelling tools and deliver an R package that will allow users to accurately estimate QALYs in a wide range of situations. We hope this will encourage a larger use of QALYs in clinical trials and better consideration of patients' preferences in medical decision-making.

36: An Alternative Estimand for Overall Survival in the Presence of Treatment Discontinuation: Simulation Results and Case Study

Kara-Louise Royle¹, David Meads², Jennifer Visser-Rogers³, David A Cairns¹, Ian R White⁴

¹Leeds Cancer Research UK Clinical Trials Unit, Leeds Institute of Clinical Trials Research, University of Leeds, Leeds, UK

²Academic Unit of Health Economics, Leeds Institute of Health Sciences, University of Leeds, Leeds, UK

³Coronado Research, Kent, England

⁴MRC Clinical Trials Unit at UCL, London, UK

Introduction Overall survival (OS) is a definitive endpoint for clinical effectiveness in cancer clinical trials. However, intercurrent events, like treatment discontinuation, can affect its interpretation.

A recent literature review concluded that treatment discontinuation and the uptake of subsequent anti-cancer treatment is often considered part of the treatment strategy i.e. researchers follow the “Treatment Policy” approach.

Our objective was to investigate the novel alternative hypothetical estimand: What is the effect on OS of the experimental trial treatment versus the control treatment, if all participants who discontinued prior to death received the same subsequent treatment?

Methods Statistical techniques, including simple intention-to-treat (ITT) and per protocol (PP) methods and more complex two-stage and inverse proportional censoring weighting (IPCW) methods, were applied in a simulation study. The data-generating mechanism simulated a two-arm randomised controlled trial dataset of 700 participants and three stratification factors. Observed and unobserved variables were simulated at baseline and follow-up timepoints. At each follow-up timepoint, some participants were simulated to discontinue and start one of two (A or B) subsequent treatments. Eleven different scenarios were considered, including varying the true experimental treatment effect and timing of treatment discontinuation. The estimand of interest was the hazard ratio and 95% confidence interval of the experimental vs control arms if everyone who discontinued had the same subsequent treatment (A rather than B). The methods were evaluated in terms of bias, coverage, and power, calculated across 1000 repetitions.

Results The ITT method was biased across all scenarios, but mostly had adequate power and coverage. The PP methods were biased with poor coverage in all scenarios. The two-stage methods were unbiased and had adequate power and coverage in almost all scenarios. The IPCW methods' performance fluctuated the most across the scenarios.

Discussion The simulation study found that the estimand could be estimated, with varying levels of performance, by all implemented methods. Overall, the two-stage method was the most consistently accurate method across the scenarios. The practicability of estimating the hypothetical estimand using the two-stage method in practice will be assessed through a real clinical trial case study, presented at the meeting. The trial was chosen as second-

line immunotherapy was introduced during trial follow-up. As more effective treatments are developed, this is likely to be a common scenario. We will discuss the generalisability of the hypothetical estimand, how it improves the interpretation of clinical trial results and the necessary considerations when analysing OS in such situations.

37: Corrections of Confidence Interval for Differences in Restricted Mean Survival Times in Clinical Trials with Small Sample Sizes

Hiroya Hashimoto¹, Akiko Kada^{2,1}

¹NHO Nagoya Medical Center, Japan

²Fujita Health University, Japan

Background / Introduction In recent years, restricted mean survival time (RMST) has been used as a measure to demonstrate the difference in efficacy between treatment groups in time-to-event outcomes, especially when the proportional hazards assumption does not hold. However, statistical tests and interval estimations based on asymptotic normality may deviate from the normal distribution when the sample size is small, leading to an inflation of the type I error rate. In this presentation, we discuss the correction of confidence intervals for between-group differences in RMST.

Methods Under the condition that the survival functions of two groups follow the same Weibull distribution and the censoring functions follow a uniform distribution, we conducted a simulation analysis in a two-group comparative study with fewer than 100 subjects per group under various scenarios. We examined the following methods:

- (1) A method based on asymptotic normality,
- (2) A method that applies bias correction to the standard error of Method (1), specifically, multiplying the standard error for each group by $\sqrt{m_i/(m_i-1)}$, where m_i is the number of events in group i ,
- (3) A method that lies between Methods (1) and (2), specifically, multiplying the standard error for each group by $\sqrt{m_i/(m_i-0.5)}$.

Results As expected, Method (1) had the highest Type I error rate in all scenarios considered, followed by Method (3), and Method (2) had the lowest. In the uncensored situation, Method (2) was generally the most appropriate, and Method (3) was optimal when the event rate

was low ($S(\tau)=0.7$). Method (2) also tended to be too conservative as the censoring rate increased, and this was more pronounced for smaller sample sizes.

Conclusions Method (3) produces better results when events occur less frequently. Method (2) yields conservative results, but caution should be exercised because it is too conservative in situations with small sample sizes and high censoring. When the sample size per group exceeds 100, the difference between methods is negligible.

Wednesday Posters at Biozentrum

Wednesday, 2025-08-27 09:00 - 10:30, Biozentrum, 2nd floor

1: Cure Models to Compare Aftercare Monitoring Schemes in Pediatric Cancer

**Ulrike Pötschger¹, Harm van Tinteren², Evgenia Glogova¹, Helga Arnardottir¹,
Paulina Kurzmann¹, Sabine Taschner-Mandl¹, Lieve Titgat², Martina Mittlböck³**

¹St. Anna Children's Cancer Research Institute, Austria

²Princess Maxima Center

³Medical University of Vienna, Center for Medical Data Science

Background / Introduction Neuroblastoma is a malignant tumor of the peripheral nervous system and 50% of the patients are high risk with a poor outcome. Monitoring with minimally invasive liquid biopsies may now allow earlier detection of tumor recurrence compared to conventional follow up evaluations based on imaging and bone marrow biopsies.

In a randomized study two monitoring strategies for relapsed Neuroblastoma are compared: minimal invasive liquid biopsies-based monitoring and conventional follow-up evaluations imaging and bone marrow biopsies.

The primary endpoint is disease-free survival (DFS). When liquid biopsies monitoring is beneficial, disease recurrences can be detected earlier. Thus, survival curves are expected to show an early group difference that vanishes in the long-term and consequently non-proportional hazards are expected.

Methods The primary statistical evaluation of the treatment effect will be done with a Weibull mixture cure model. The crucial assumption underlying a mixture Cure model is that DFS results from the survival experience of two subgroups: cured patients and uncured patients. Within this model the proportion of cured patients and the time of an event for the uncured subpopulation are modelled separately. The time to detect a recurrence in the subpopulation of uncured patients is of primary interest here.

Monte-Carlo simulations were performed to evaluate power and statistical properties of the Weibull mixture cure-model. For the standard monitoring arm, the inversion method is used to simulate survival data following a mixture cure model as observed in historical populations. For the experimental arm the effect of different data-generating processes and liquid biopsy schedules are explored.

Results Simulation studies helped to explore different liquid biopsy schedules, effect sizes (lag time between detectable signals with liquid biopsy and imaging) under various data generating processes. Accordingly, the simulation studies helped to refine the study-design and schedule of the liquid biopsies. As compared to a conventional analysis with a Cox regression model, substantial gains in statistical power could be achieved. With a two-sided alpha of 5% and n=150 patients, the simulated power to detect recurrences 5 months earlier was 81% and 60% for the Cure- and Cox-model, respectively.

Conclusion Comparing aftercare evaluations with different schedules and sensitivities is methodologically challenging. With anticipated non-proportional hazards, it is important to directly address the primary interest in an earlier signal-detection. Simulation studies helped to assess power and to develop an optimal monitoring schedule. Cure-models provide results with a clear interpretation and lead to substantial gains in statistical power.

2: Comparison of Treatment Sequences in Advanced Pancreatic Cancer

Norbert Marschner^{1,2}, Nina Haug³, Susanna Hegewisch-Becker⁴, Marcel Reiser⁵, Steffen Dörfel⁶, Rüdiger Liersch⁷, Hartmut Linde⁸, Thomas Wolf⁹, Anna Hof¹⁰, Anja Kaiser-Osterhues², Karin Potthoff², Martina Jänicke¹⁰

¹Med. Klinik 1, Universitätsklinik Freiburg, Freiburg, Germany

²Medical Department, iOMEDICO, Freiburg, Germany

³Biostatistics, iOMEDICO, Freiburg, Germany

⁴Hämatologisch-Onkologische Praxis Eppendorf (HOPE), Hamburg, Germany.

⁵PIOH-Praxis Internistische Onkologie und Hämatologie, Köln, Germany

⁶Onkozentrum Dresden/Freiberg, Dresden, Germany

⁷Hämatologisch-onkologische Gemeinschaftspraxis, Münster, Germany

⁸MVZ für Blut- und Krebserkrankungen, Potsdam, Germany

⁹BAG, Gemeinschaftspraxis Hämatologie-Onkologie, Dresden, Germany

¹⁰Clinical Epidemiology and Health Economics, iOMEDICO, Freiburg, Germany

There are no clear guidelines regarding the optimal treatment sequence for advanced pancreatic cancer, as head-to-head phase III randomised trials are missing. We assessed real-world effectiveness of three frequently administered sequential treatment strategies: FOLFIRI-NOX→GEMNAB, GEMNAB→FOLFOX/OFF and GEMNAB→ NALIRI + 5-FU. To this end, we emulated a hypothetical target trial where patients were randomised to one of these sequences before the beginning of first-line therapy. As causal estimand, we quantified the per-protocol effect of treatment on overall survival and time-to-deterioration of health-related quality of life. Treatment effects were estimated both for the whole population and strat-

ified by risk group according to the Pancreatic Cancer Score¹. Our analysis included 1551 patients with advanced pancreatic cancer from the prospective, clinical cohort study Tumour Registry Pancreatic Cancer receiving FOLFIRINOX ($n = 613$) or gemcitabine/nab-paclitaxel (GEMNAB; $n = 938$) as palliative first-line treatment. We used marginal structural modeling to adjust for time-varying confounding affecting the relation between treatment and endpoint — a key challenge in real-world data analysis². The estimated effectiveness of the three treatment sequences evaluated was largely comparable. Patients with poor prognosis might benefit from intensified treatment with FOLFIRINOX→GEMNAB in terms of survival and quality of life. Future randomised trials on sequential treatments in advanced pancreatic cancer are warranted.³

1. Marschner N, Hegewisch-Becker S, Reiser M, von der Heyde E, Bertram M, Hollerbach SH, Kreher S, Wolf T, Binninger A, Chiabudini M, Kaiser-Osterhues A, Jänicke M, et al. FOLFIRINOX or gemcitabine/nab-paclitaxel in advanced pancreatic adenocarcinoma: A novel validated prognostic score to facilitate treatment decision-making in real-world. *Int J Cancer* 2023;152:458–69.
2. Robins JM, Hernán MA, Brumback B. Marginal Structural Models and Causal Inference in Epidemiology. In: Epidemiology. 2000. 550–60.
3. Marschner N, Haug N, Hegewisch-Becker S, Reiser M, Dörfel S, Lerchenmüller C, Linde H, Wolf T, Hof A, Kaiser-Osterhues A, Potthoff K, Jänicke M, et al. Head-to-head comparison of treatment sequences in advanced pancreatic cancer—Real-world data from the prospective German TPK clinical cohort study. *Intl Journal of Cancer* 2024;155:1629–40.

3: Clinical Trials with Time-to-Event-Endpoint: Interim Prediction of Number of Events with Confidence Distributions

Edoardo Ratti, Maria Grazia Valsecchi, Stefania Galimberti

Bicocca Bioinformatics Biostatistics and Bioimaging B4 Center, School of Medicine and Surgery, University of Milan-Bicocca, Monza, Italy

Introduction An important aspect in randomized clinical trials design is planning interim analyses. With time-to-event endpoints, the target sample size is function of event number. It is crucial that studies provide sufficient follow-up to observe the event number needed to preserve power. Novel approaches were developed in a bayesian framework to predict the date at which the target number of events is reached (maximum information trial). However, there is little on forecasting number of events expected at a fixed future date with corresponding

prediction interval in trials with fixed follow-up time (maximum duration trial).

Methods Based on a recent paper on the use of confidence distributions in clinical trials [1], we adapt a prediction method developed in reliability analysis [2] and show its potential in the clinical context. The proposed method obtains prediction intervals from a predictive distribution constructed on a bootstrap-based confidence distribution of the parameters of the fitted survival model. The appropriateness of the framework was assessed by application on a real phase III trial and by evaluating intervals coverage probability with simulations.

Results Using data from a published phase III trial [3], at the second interim (accrual closed) and every subsequent 6 months we predicted the number of occurring event. Results show that all intervals included the observed number of events. Simulations show that prediction intervals have the desired coverage under the appropriate survival distribution.

Conclusions For a maximum duration trial, it is crucial to predict the number of events at future times with proper prediction intervals. The presented approach allows to construct valid predictive inference based on confidence distributions accommodating different parametric model/censoring mechanisms. This is an alternative to a bayesian approach. Its use is proposed here for prediction after accrual closure and further work will face modelling accrual.

References

- [1] Marschner IC., Confidence distributions for treatment effects in clinical trials: Posteriors without priors. *Statistics in Medicine*. 2024; 43(6): 1271-1289
- [2] Tian 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
- [3] Conter V, Valsecchi MG, Cario G, et al. Four Additional Doses of PEG-L-Asparaginase During the Consolidation Phase in the AIEOP-BFM ALL 2009 Protocol Do Not Improve Outcome and Increase Toxicity in High-Risk ALL: Results of a Randomized Study. *J Clin Oncol*. 2024 Mar 10;42(8):915-926.

4: A Bayesian-Informed Dose-Escalation Design for Multi-Cohort Oncology Trials with Varying Maximum Tolerated Doses

Martin Kappler¹, Yuan Ji²

¹Cytel Inc., Waltham, USA

²University of Chicago, USA

In oncology dose-escalation trials, it is common to evaluate a drug across multiple cancer types within the same study. However, different cancer types may also have different maximum tolerated doses (MTDs) due to potentially different underlying patient characteristics. Standard approaches either pool all patients, potentially ignoring important differences between cancer types, or conduct separate dose-escalation processes for each type, which can lead to inefficiencies. We propose a dose-escalation design that leverages the dose-level information from faster-recruiting cohorts to inform dose-escalation and de-escalation rules for slower-recruiting cohorts, thereby balancing safety, efficiency, and cohort-specific MTD estimation.

Our approach is based on a model assisted dose escalation design and uses informative priors to leverage dose-toxicity information from the faster-recruiting cohort to the slower-recruiting cohort. This approach enables a more conservative and adaptive dose-escalation process for slower cohorts by updating the prior based on observed dose-limiting toxicities in the faster cohort. The informative prior ensures that the dose-escalation in the slower cohort is both cautious and responsive to emerging data, without requiring separate dose-escalation processes for each cancer type. Uncertainty for slower cohorts is reduced and unnecessary toxicity risks are avoided.

The operating characteristics of the approach (probability to determine MTD, number of patients exposed to toxic doses, etc.) are assessed via simulations over a variety of scenarios in the two cohorts and are compared to separate or pooled escalation.

5: Comparison of Bayesian Approaches in Single-Agent Dose-Finding Studies

Vibha Srichand

Prasanna School of Public Health, Manipal Academy of Higher Education, India

Single-agent dose-finding studies conducted as part of phase 1 clinical trials aim to obtain sufficient information regarding the safety and tolerability of a drug, with the primary objective of determining the Maximum Tolerated Dose (MTD) – the maximum test dose that can be administered with an acceptable level of toxicity. While the 3+3 design has been the conventional choice for dose-finding studies, innovative Bayesian designs have gained prominence. These designs provide a framework to incorporate prior knowledge with data accumulated during the study to adapt the study design and efficiently estimate the MTD. However, existing Bayesian assume a specific parametric model for the dose-toxicity relationship which reduces its adaptability to complex data patterns. To address this limitation, recent research has introduced nonparametric Bayesian methods which are model free, robust

and well-suited for small sample sizes. Thus, it is imperative to comprehensively compare the performance of parametric and nonparametric Bayesian methods and provide evidence for the implementation of different methods.

This paper aims to understand the accuracy, safety and adaptability of dose-finding methods by analysing different scenarios of target toxicity probabilities and varying cohort sizes for a predetermined sample size. The methods under review are as follows: traditional method – 3+3 design; parametric methods – continual reassessment method (CRM), modified toxicity probability (mTPI and mTPI-2), keyboard and Bayesian optimal interval designs (Kurzrock et al., 2021) as well as nonparametric methods – Bayesian nonparametric continual reassessment (Tang et al., 2018) and Bayesian stochastic approximation method (Xu et al., 2022). The performance of the designs will be assessed using four key metrics, with conclusions drawn based on extensive simulation studies.

Keywords Dose-finding, Maximum Tolerated Dose, Clinical trial design, Bayesian, Parametric, Nonparametric, Continual Reassessment method, Stochastic Approximation

References

Kurzrock, R., Lin, C.-C., Wu, T.-C., Hobbs, B. P., Pestana, R. C., MD, & Hong, D. S. (2021). Moving beyond 3+3: The future of clinical trial design. American Society of Clinical Oncology Educational Book. American Society of Clinical Oncology. Meeting, 41, e133–e144. https://doi.org/10.1200/EDBK_319783

Tang, N., Wang, S., & Ye, G. (2018). A nonparametric Bayesian continual reassessment method in single-agent dose-finding studies. BMC Medical Research Methodology, 18(1), 172. <https://doi.org/10.1186/s12874-018-0604-9>

Xu, J., Zhang, D., & Mu, R. (2022). A dose-finding design for phase I clinical trials based on Bayesian stochastic approximation. BMC Medical Research Methodology, 22(1), 258. <https://doi.org/10.1186/s12874-022-01741-3>

6: Evaluating the Effect of Different Non-Informative Prior Specifications on the Bayesian Proportional Odds Model in Randomised Controlled Trials

Chris J Selman^{1,2}, Katherine J Lee^{1,2}, Michael Dymock^{3,4}, Ian Marschner⁵, Steven Y.C. Tong^{6,7}, Mark Jones^{4,8}, Tom Snelling^{3,8}, Robert K Mahar^{1,9,10}

¹Clinical Epidemiology and Biostatistics Unit, Murdoch Children's Research Institute, Australia

²Department of Paediatrics, University of Melbourne, Australia

³Wesfarmers Centre of Vaccines and Infectious Diseases, The Kids Research Institute Australia, Australia

⁴School of Population and Global Health, The University of Western Australia, Australia

⁵NHMRC Clinical Trials Centre, Faculty of Medicine and Health, The University of Sydney, Sydney, NSW 2050, Australia

⁶Victorian Infectious Diseases Services, The Royal Melbourne Hospital, Australia

⁷Department of Infectious Diseases, University of Melbourne, Australia

⁸Sydney School of Public Health, Faculty of Medicine and Health, University of Sydney, Australia

⁹Centre for Epidemiology and Biostatistics, University of Melbourne, Australia

¹⁰Methods and Implementation Support for Clinical and Health Research Hub, University of Melbourne, Australia

Background Ordinal outcomes can be a powerful way of combining multiple distinct patient outcomes into a single endpoint in randomised controlled trial (RCT). Such outcomes are commonly analysed using proportional odds (PO) models. When the analysis uses a Bayesian approach, it is not obvious what ‘non-informative’ priors should be used and whether these are truly ‘non-informative’, particularly in adaptive trials where early stopping decisions may be influenced by the choice of prior.

Methods This study evaluates the effect of different non-informative prior specifications on the Bayesian PO model for a two-arm trial in the context of a design with an early stopping rule and a fixed design scenario. We conducted an extensive simulation study, varying factors such as effect size, sample sizes, number of categories and the distribution of the control arm probabilities. The models are also illustrated using data from the Australian COVID-19 Trial.

Results Our findings indicate that the prior specification can introduce bias in the estimation of the treatment effect, particularly when control arm probabilities are right-skewed. Using an R-square prior specification had the smallest bias and increased the likelihood of stopping early in such settings when there was a treatment effect. However, this specification exhibited larger biases for control arm probabilities that were U-shaped and trials that incorporated an early stopping rule. Using Dirichlet priors with concentration parameters close to zero had the smallest bias when probabilities were right-skewed in the control arm, and were more likely to stop earlier for superiority for trials that incorporated early stopping rules even if there was no treatment effect. Specifying concentration parameters close to zero using the Dirichlet prior may also cause computational issues at interim analyses with small sample sizes and larger number of categories in the outcome.

Conclusion The specification of non-informative priors in Bayesian adaptive trials that use ordinal outcomes has implications for treatment effect estimation and early stopping decisions.

Careful selection of priors that consider the likely distribution of control arm probabilities or informed sensitivity analyses may be essential to inference is not unduly influenced by inappropriate priors.

7: Bayesian Decision Analysis for Clinical Trial Design with Binary Outcome in the Context of Ebola Virus Disease Outbreak – Simulation Study

Drifa Belhadi^{1,2}, Joonhyuk Cho^{3,4,5}, Pauline Manchon⁶, Denis Malvy^{7,8}, France Mentré^{1,6}, Andrew W Lo^{3,5,9,10}, Cédric Laouénan^{1,6}

¹Université Paris Cité, Inserm, IAME, F-75018 Paris, France

²Saryga, France

³MIT Laboratory for Financial Engineering, Cambridge, MA, USA

⁴MIT Department of Electrical Engineering and Computer Science, Cambridge, MA, USA

⁵MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, USA

⁶AP-HP, Hôpital Bichat, Département d'Épidémiologie Biostatistique et Recherche Clinique, F-75018 Paris, France

⁷UMR 1219 Inserm/EMR 271 IRD, University of Bordeaux, Bordeaux, France

⁸Department for Infectious and Tropical Diseases, University Hospital Center Pellegrin, Bordeaux, France

⁹MIT Operations Research Center, Cambridge, MA, USA

¹⁰MIT Sloan School of Management, Cambridge, MA, USA

Background When designing trials for high-mortality diseases with limited available therapies, the conventional 5% type I error rate used for sample size calculation can be questioned. Bayesian Decision Analysis (BDA) for trial design allows for the integration of multiple health consequences of the disease when designing trials. This study adapts BDA for trials with binary outcomes to calculate optimal sample sizes and type I error rates in the context of an Ebola virus disease outbreak.

Methods We consider a fixed, two-arm randomized trial with a binary outcome and two types of clinical trial loss: post-trial loss, for not approving an effective treatment or approving an ineffective treatment; in-trial loss, for not administrating an effective treatment to patients in the control arm or for administrating an ineffective treatment for patients in the experimental arm. The model accounts for side effects of an ineffective treatment and the burden of Ebola disease. A loss function was defined to summarize the multiple consequences into a single measure, and optimal sample sizes (n) and type I error rates (α) were derived by minimizing this loss function.

Results Using the mortality rate as the outcome, we varied model parameters to represent different Ebola epidemic scenarios, such as target population size, mortality rate, and treatment efficacy. In most cases, BDA-optimal α values exceeded the conventional one-sided 2.5% rate and BDA-optimal sample sizes were smaller. Additionally, we conducted simulations comparing a BDA-optimized two-arm trial (fixed or sequential) to standard designs (two-arm/single-arm, fixed/sequential) across various outbreak scenarios. Overall, statistical power remained comparable across designs, except when sample size assumptions were incorrect, or when the trial started after the outbreak peak; in these situations, BDA-optimized trials were associated with superior powers.

Conclusion This BDA adaptation provides a new framework for designing trials with a binary outcome, enabling more effective evaluation of therapeutic options. It is particularly valuable for diseases with high mortality rates and limited treatment options. In an outbreak context, where case numbers decline after the epidemic peak and there is uncertainty around mortality rate and treatment efficacy, BDA-optimized trials offer an interesting approach for evaluating new experimental treatments.

8: Relevance of Electronic Medical Records for Clinical Trial Eligibility: A Feasibility Assessment in Acute Stroke Studies

Yusuke Sasahara¹, Taizo Murata², Yasufumi Gon^{3,4}, Toshihiro Takeda^{2,5}, Eisuke Hida¹

¹Department of Biostatistics and Data Science, Osaka University Graduate School of Medicine

²Department of Medical Informatics, Osaka University Hospital

³Department of Neurology, Osaka University Graduate School of Medicine

⁴Academic Clinical Research Center, Osaka University Hospital

⁵Department of Integrated Medicine, Medical Informatics, Osaka University Graduate School of Medicine

Electronic medical records (EMRs) are a key source of real-world data in clinical trials. In hyperacute-phase diseases, where conducting RCTs is challenging, external control arms using EMRs are expected to enhance trial feasibility. In July 2024, FDA released guidance on evaluating EMRs and claims data to support regulatory decision-making, emphasizing the importance of ensuring data reliability and relevance. However, evidence on how well EMRs meet these criteria remains limited. This study evaluates the feasibility of extracting clinical trial eligibility criteria from EMRs, focusing on data extraction and structuring in acute stroke studies.

Abstracts of Contributed Posters

Five acute stroke-related clinical trials with detailed eligibility criteria were selected from the jRCT and UMIN-CTR databases. Registration forms were created for each trial, and an expert panel (physician, medical informatician, statistician, and data manager) evaluated the feasibility of extracting these criteria from EMRs at Osaka University Hospital. Data types were categorized into four groups: structured, mosaic (a mix of structured and unstructured), unstructured, and unavailable. The proportion of each type was summarized by trial and item category, and extraction feasibility was scored (structured: 3, mosaic: 2, unstructured: 1, unavailable: 0). Data were visualized using bar charts, box plots, and radar charts.

Across all five trials, structured data accounted for 37.6%, mosaic for 12.1%, unstructured for 42.3%, and unavailable for 8.1%. The proportion of unstructured data varied among trials, with Trial B having the highest (68.3%) and Trial C the lowest (15.8%). Trial A had the highest unavailable data (16.7%). Imaging-related variables were entirely unstructured (100%), and medical history/comorbidity (84.6%) and diagnosis (61.1%) also lacked structure. In contrast, structured data were demography (80.0%), treatment applicability (62.5%), and laboratory/vital signs (56.3%).

The study assessed how well EMRs align with clinical trial eligibility criteria to evaluate their relevance. Due to the variability in EMRs availability across trials and items, a preliminary assessment is necessary for each protocol. Since 42.3% of all items were unstructured, manual chart review may be unavoidable. Structured data were more prevalent in demography and treatment applicability, whereas imaging and medical history/comorbidity data posed major challenges. FDA guidelines highlight the need for validation and bias assessment in data transformation, requiring standardized processes to enhance EMR relevance for regulatory use.

The feasibility of extracting eligibility criteria and the degree of structuring in EMRs varied across trials and items. While imaging and medical history/comorbidity data were poorly structured, developing standardized data extraction methods may enhance the relevance of EMRs.

9: Navigating Complex and Computationally Demanding Clinical Trial Simulation

Saumil Shah, Mitchell Thomann

Boehringer Ingelheim, Germany

Many diseases lacking treatment options have multiple correlated endpoints as progression biomarkers. Establishing efficacy in many endpoints with randomised dose-finding represents

an unmet need.

A seamless Phase IIa-IIb trial design was proposed, featuring staggered recruitment, dropouts, longitudinal and correlated endpoints, and interim analysis. The trial design also included using historical information using Bayesian meta-analytic priors and Bayesian dose-finding methods to improve trial efficacy. Scenario planning across a wide range of effects, dose-response models and endpoint correlations is a considerable challenge. Thus, a robust trial simulation implementation was required to estimate operating characteristics precisely and optimise the study design.

We used the random slope and intercept method to capture the longitudinal endpoint and patient-level variance. The correlated secondary endpoint was generated from conditional distributions. The informative historical prior was updated with the generated data to get a posterior. We used the posterior in the interim analysis to compare the across-arm gains in the change from baseline values. The final analysis used the multiple comparison procedure and Bayesian modelling for randomised dose-finding. We considered six appropriate candidate dose-response models for the Bayesian modelling. Each endpoint was assigned go-no-go boundaries for stop, continue or success decisions. We used the median of posterior distribution from the fitted Bayesian models to make the decision.

We used R programming language and available open-source packages to implement the trial simulation. The data generation and analysis steps were implemented as a collection of functions in a pipeline. The pipeline was managed using {targets}, a workflow management package. Such management allowed us to handle many scenarios and replicates, preventing redundant and unnecessary computations. It also helped with parallel execution, bringing execution time to the order of hours on a high-performance cluster.

Our implementation enabled the rapid exploration of a wide range of trial scenarios and treatment effects, enabling reliable estimation of the operating characteristics of each design aspect. This approach provides a potent tool for optimising clinical trial design across therapeutic areas.

10: Transforming Clinical Trials: The Power of Synthetic Data in Augmenting Control Arms

Emmanuelle Boutmy¹, Shane O Meachair², Julie Zhang⁵, Sabrina de Souza¹, Saheli Das⁴, Dina Oksen³, Anna Tafuri¹, Lucy Mosquera^{5,6}

¹Merck KGaA, Darsmtadt, Germany

²Action, Barcelona, Spain

³Merck Biopharma Co., Ltd.

⁴Merck Specialities Pvt., Ltd.

⁵Action, New York, USA

⁶CHEO Research Institute, Ottawa, Ontario, Canada

Background Synthetic data generation (SDG) creates artificial datasets that replicate the characteristics of clinical trial (CT) data, potentially mitigating challenges when real data is scarce. This study aimed to explore methods for Synthetic Data Augmentation (SDA): augmentation (adding synthetic data to original data) of a CT control arm. Data from the INTRAPID lung 0037 control arm were used, consisting of advanced NSCLC patients with high PD-L1 expression treated with pembrolizumab (n=152).

Methods Three generative models were employed to create synthetic data: Sequential decision trees (SDT), Bayesian networks (BN), and Transformer synthesis (TS), alongside a reference approach using bootstrapping (BS). Descriptive statistics, parameter estimates, and standard errors were calculated using a multiple imputation method for synthetic data using 10 synthetic datasets. The quality of synthetic data was assessed through utility and workload-specific assessments. The primary outcome was bias relative to the full CT control arm estimate and standard deviation to assess variability across samples. Bias assessments compared augmented estimates to Progression Free Survival (PFS) from the full control arm, simulating scenarios with 50% unavailable control arm data.

Results Univariate distances and multivariate relationships were below the pre-specified threshold indicating close replication of real data distribution for all models except TS. Results indicated that synthetic data produced outcomes comparable to real data, with bias in PFS ranging from -2.69 for TS to +0.2 months for SDT, where values closer to zero indicates better performance (-2.22 for BN, -0.71 months for BS). SDT synthesis demonstrated the lowest bias among all augmentation methods including the reduced control sample alone. In the sensitivity analysis, SDT was the only approach whose 95% interval included the true ground truth PFS estimate from the full CT control arm.

Discussion These results suggest that generative models could yield nearly identical distributions for real and synthetic variables. SDA has been shown to yield estimates with low bias compared to using the available CT data alone and may be leveraged for clinical trials where patient enrollment in the control arm is difficult, such as simulating trial scenarios or completing datasets for underrepresented groups. Further research is needed to confirm and develop a synthetic validation framework to assess the limits of SDA for statistical inference, as well as impact on other statistical quantities such as power and type I error rates and harness the transformative power of synthetic data.

11: Integrating Stakeholder Perspectives in Modeling Routine Data for Therapeutic Decision-Making

Michelle Pfaffenlehner^{1,2}, Andrea Dreßing^{3,4}, Dietrich Knoerzer⁵, Markus Wagner⁶, Peter Heuschmann^{7,8,9}, André Scherag¹⁰, Harald Binder^{1,2}, Nadine Binder^{2,11}

¹Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center – University of Freiburg, Germany

²Freiburg Center for Data Analysis and Modeling and AI, University of Freiburg, Freiburg, Germany

³Department of Neurology and Clinical Neuroscience, Medical Center, University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany

⁴Freiburg Brain Imaging Center, Faculty of Medicine, Medical Center–University of Freiburg, University of Freiburg, Freiburg, Germany

⁵Roche Pharma AG, Grenzach, Germany

⁶Stiftung Deutsche Schlaganfall-Hilfe, Gütersloh, Germany

⁷Institute for Medical Data Sciences, University Hospital Würzburg, Würzburg, Germany

⁸Institute for Clinical Epidemiology and Biometry, University Würzburg, Würzburg, Germany

⁹Clinical Trial Centre, University Hospital Würzburg, Würzburg, Germany

¹⁰Institute of Medical Statistics, Computer and Data Sciences, Jena University Hospital – Friedrich Schiller University Jena, Jena, Germany

¹¹Institute of General Practice/Family Medicine, Faculty of Medicine and Medical Center - University of Freiburg, Freiburg, Germany

Background Routine medical data offer a valuable resource for generating evidence to improve patient care in therapeutic contexts beyond randomized controlled trials. These data include patient-related parameters, diagnostic information, and treatment data recorded in digital patient records from hospital admission to discharge. With the introduction of the German Health Data Use Act (GDNG) in 2024, the use of such data is becoming more accessible in Germany. However, methodological approaches must account for the diverse needs of stakeholders, including clinicians, the pharmaceutical industry, patient advocacy groups, and statistical modelers. This study explores how different perspectives shape the use and interpretation of routine data in medical decision-making, with each perspective aiming to address specific research questions.

Methods Building on insights from an interdisciplinary workshop that we recently organized, we examine how various stakeholder perspectives can be incorporated into the modelling of routine data. We discuss key routine data sources, such as electronic health records, and highlight statistical and artificial intelligence (AI)-based techniques that could be used to extract meaningful insights. Moreover, the linkage of patient-reported outcomes will be discussed to address the patient's perspective. Additionally, we illustrate how different modelling

approaches address distinct research questions, reflecting the priorities of the stakeholder groups. A particular focus is placed on multi-state models, which are well-suited for capturing disease and treatment trajectories by structuring diagnoses and treatments as transition events over time.

Results Our conceptual analysis identifies multiple approaches for integrating diverse perspectives into routine data modelling. For example, clinicians prioritize clinical relevance and interpretability, the pharmaceutical industry focuses on regulatory compliance and real-world evidence, while patient representatives emphasize transparency and inclusion of patient-reported outcomes. Multi-state models are particularly advantageous because they allow the characterization of dynamic disease processes and patient transitions between states, offering a more accessible and interpretable approach to routine data analysis. Still, challenges remain in data quality.

Conclusion Effective use of routine data in medical decision-making requires robust analytical methods that meet the needs of diverse stakeholders. Multi-state models provide a dynamic framework for capturing disease progression and treatment pathways, making them particularly suitable for clinical and regulatory applications. To maximize their impact, future research should focus on improving data integration, transparency in methods used, and making the methods practically useful, leading to better integration into healthcare decision-making.

12: Aligning Synthetic Trajectories from Expert-Based Models with Real Patient Data Using Low-Dimensional Representations

Hanning Yang¹, Meropi Karakioulaki², Cristina Has², Moritz Hess¹, Harald Binder¹

¹Institute of Medical Biometry and Statistics (IMBI), University of Freiburg, Germany

²Department of Dermatology, University of Freiburg, Germany

Background Quantitative models, such as ordinary differential equations (ODEs), are widely used to model dynamic processes, such as disease progression - e.g., for subsequently generating synthetic data. However, calibrating them with real patient data, which is typically sparse, noisy, and highly heterogeneous can be challenging. This is particularly notable in rare diseases like Epidermolysis Bullosa (EB), where observations are limited, and data is often missing. To address this, we developed an approach to calibrate ODEs informed by expert knowledge with real, observational patient data using low-dimensional representations.

Methods We developed an ODE system informed by experts to model EB key biomarker

dynamics and employed an autoencoder for dimensionality reduction. Calibration of ODE parameters was informed by a loss computed from the distance between real and ODE-derived synthetic observations in the latent space. Specifically, this loss captures key trajectory features, including temporal alignment and pointwise differences. To handle discrepancies in initial conditions, a centring approach is applied during early iterations, and an imputation layer is trained to address missing data.

Results A simulation study demonstrated robustness under high noise and complex missing patterns, with parameters converging to the ground truth. When applied to real EB data, our method consistently improved the alignment between synthetic and real data, despite the challenges of noisy and sparse observations from only 21 highly diverse patients. As a result, relationships in synthetic data became more consistent with real patient data.

Conclusion This study presents a novel approach for calibrating an expert-informed synthetic data model using neural networks, supporting realistic synthetic individual patient data (IPD) generation and advancing rare disease research.

13: Integrating Semantic Information in Care Pathway Studies with Medical Code Embeddings, Application to the Case of Amyotrophic Lateral Sclerosis

Corentin Faujour^{1,2}, Stéphane Bouée¹, Corinne Emery¹, Anne-Sophie Jannot^{2,3}

¹CEMKA, Bourg-La-Reine, France

²Université Paris Cité, Inria, Inserm, HeKA, F-75015 Paris, France

³French National Rare Disease Registry (BNDMR), Greater Paris University Hospitals (AP-HP), Université Paris Cité, Paris, France

Background Modelling care pathways from claims databases is a challenging task given the tens of thousands of existing medical codes, sometimes associated with the same medical concept. In such modelling, medical codes are usually represented as sets of binary variables (one-hot encoding), which does not allow for the inclusion of semantic information. Embedding medical codes in a continuous space, so that semantically related codes are represented by numerically similar vectors, could improve care pathway modelling.

We aimed to embed codes from the International Classification of Diseases (ICD-10) and the Anatomical Therapeutic Chemical Classification (ATC) into a common latent space. A secondary goal was to use these embeddings in the prediction of amyotrophic lateral sclerosis (ALS).

Methods A co-occurrence matrix between codes was constructed from care sequences contained in the ESND, a French claims database containing care consumptions for a representative sample of 1.5 million patients over 15 years. Code embeddings for all 5 classifications systems available in the ESND, i.e. representative numerical vectors that capture semantic relationships, were then obtained using singular value decomposition on the corresponding pointwise mutual information matrix.

Embeddings' consistency was assessed using UMAP visualisation and nearest neighbour searches. The resulting embeddings were used to predict the occurrence of ALS in a personalized logistic regression model, taking as input all codes in the care sequence prior to diagnosis. Sequence-level embeddings were obtained by an average-pooling operation at the code level. We compared the performance obtained using embeddings as input with those obtained using one-hot encoding.

Results We obtained embeddings for 30,000 codes, including 9,900 ICD-10 codes and 1,400 ATC codes from 1.5 million care pathways representing 400 million tokens. Consistency evaluation revealed that semantically related codes form clusters in the latent space, e.g., the diagnosis code for motor neuron disease is surrounded by other muscle disorders (myopathies, muscular dystrophy, etc.) and its specific treatment (riluzole).

Using the resulting embeddings to classify sequences from 22,000 ALS patients and 22,000 matched controls, we were able to significantly improve predictive performance (AUC: 0.78, 95% CI [0.77-0.79] with embeddings vs. 0.74 [0.73-0.75] with one-hot encoding). This suggests that the inclusion of semantic information is relevant for such a prediction task.

Conclusion This is the first semantic representation of ICD-10 and ATC codes in a common latent space, two classifications commonly used in claims databases. The resulting embeddings can be used to improve the representation of healthcare pathways.

14: Inequalities in Impact of Respiratory Viruses: Development and Analysis of Respiratory Virus Phenotypes in EHRs from England using OpenSAFELY

Em Prestige¹, Jennifer K. Quint², Charlotte Warren-Gash¹, William Hulme³, Edward PK Parker¹, Elizabeth Williamson¹, Rosalind M. Eggo¹

¹London School of Hygiene & Tropical Medicine, United Kingdom

²Imperial College London, United Kingdom

³Bennett Institute for Applied Data Science, Nuffield Department of Primary Care Health Sciences, University of Oxford, United Kingdom

Background Respiratory virus burden is large and unequally distributed in England, with disproportionate impact in socioeconomically deprived areas and minority ethnic groups. To explore these disparities using electronic health records (EHRs) computable phenotypes must be designed to identify reported respiratory virus health events. However, many EHR codes are non-specific or uncertain, for example, a patient could have codes for 'cough' or 'suspected influenza' and neither of these would be highly specific identifiers of flu cases. Therefore, sensitivity and specificity of the phenotypes should determine what codes to include. This research explores the design of phenotypes to identify patients with respiratory viruses - respiratory syncytial virus (RSV), influenza (flu), and COVID-19, and the subsequent application exploring disparities in the impact of these conditions. We highlight the trade-offs between sensitivity and specificity in phenotype design and their implications for identifying health disparities.

Methods With the approval of NHS England, we used pseudonymized GP data in Open-SAFELY, linked with Hospital Episode Statistics (HES) and ONS mortality data, to develop phenotypes for mild (primary/emergency care) and severe (secondary care) respiratory outcomes. For each virus, we created maximally sensitive and specific phenotypes to capture cases with more frequency or accuracy respectively. Maximally sensitive phenotypes included non-specific symptoms and suspected diagnosis codes, whereas, maximally specific phenotypes included lab test results. We then identified disparities by socioeconomic status and ethnicity in these outcomes from 2016-2024. We used Poisson regression for rates of mild and severe outcomes per 1000 person-years, adjusting for age group, sex, rurality, and where relevant, vaccination status. We performed analyses on the NHS records of approximately 45% of England's population, presenting a unique opportunity to explore respiratory outcomes in cohorts where cases are rare or under-ascertained.

Results We report differences and overlap in cases identified using specific versus sensitive phenotypes across the three pathogens. We describe the extent to which disparities in respiratory outcomes vary by pathogen, age cohort and severity of disease and use adjusted models to explore patterns of risk across ethnicity and socioeconomic status in different phenotypes.

Conclusion Both highly specific and sensitive computable phenotypes are essential tools in EHR research. Their design should align with research objectives, balancing accuracy with the required number of outcomes. Exploring multiple phenotype definitions supports sensitivity analyses and subgroup evaluations. Furthermore, disparities in respiratory virus outcomes highlight the pathogen-specific risks and age-related vulnerabilities that should be targeted to minimise health inequities.

15: Modeling Longitudinal Clinical Outcomes: Comparison of Generalized Linear Models, Generalized Estimating Equations, and Marginalized Multilevel Models in Pediatric Intensive Care

Luca Vedovelli¹, Stefania Lando¹, Danila Azzolina², Corrado Lanera¹, Ileana Baldi¹, Dario Gregori¹

¹University of Padova, Italy

²University of Ferrara, Italy

Introduction Longitudinal data analysis is essential in neonatal and pediatric intensive care, where patient outcomes evolve rapidly, such as in sepsis progression or respiratory distress. Selecting the right statistical model is critical for accurate clinical effect estimation. We compared four modeling approaches—generalized linear models (GLM), GLM with a shrinkage factor, generalized estimating equations (GEE), and marginalized multilevel models (MMM)—in scenarios replicating real-world complexity, including random effects, latent effects, and transition dynamics. Our study evaluated model accuracy, robustness, and interpretability in small and variable cluster settings typical of intensive care units, where patient populations are often limited, heterogeneous, and subject to rapid physiological changes.

Methods We conducted a simulation study reflecting the heterogeneity of clinical trajectories in neonatal and pediatric intensive care. Scenarios included non-fixed patient clusters ranging from 4 to 10 and sample sizes between 20 and 150. Models were evaluated based on Mean Absolute Percentage Error (MAPE), Type I and Type II error rates, and parameter stability. We assessed the impact of incorporating shrinkage factors in GLM to mitigate estimation biases.

Results MMM consistently outperformed GEE and GLM in small sample sizes and low cluster counts, yielding lower MAPE and reduced bias. This superior performance is due to its integration of marginal and subject-specific effects while accounting for within-cluster correlation. As sample size and cluster numbers increased, performance differences diminished. GEE and GLM exhibited high variability in small samples, with GEE particularly unstable. GLM tended to overestimate effects, inflating Type I error rates. MMM maintained a controlled Type I error rate, though at the cost of slightly reduced power.

Conclusion In neonatal and pediatric intensive care, where patient populations are small and heterogeneous, MMM is a more reliable alternative to GEE and GLM. It balances interpretability and robustness, making it well suited for longitudinal clinical applications. While GLM is adequate in large datasets, its tendency to overestimate effects warrants caution, as it may misguide clinical decisions. GEE, although widely used, is less stable in small samples. Our findings support the use of MMM for clinical research requiring accurate inference

of treatment effects and patient trajectories. Future work should explore Bayesian extensions of MMM for enhanced inferential precision through improved uncertainty modeling, small-sample estimation, and incorporation of prior knowledge.

16: Modelling the Costeffectiveness of Truvada for the Prevention of Mother to Child Prevention (PMTCT) of Hepatitis B Virus in Botswana

Graceful Mulenga^{1,2}, Motswedi Anderson^{1,4,5}, Simani Gaseitsiwe^{1,3}

¹Botswana Harvard Health Partnership, Botswana

²Department of Mathematics and Statistical Sciences, Faculty of Science, Botswana International University of Science and Technology, Palapye , Botswana

³Department of Immunology and Infectious Diseases, Harvard T. H. Chan School of Public Health, Boston, MA, USA

⁴The Francis Crick Institute, London, UK

⁵Africa Health Research Institute, Durban, South Africa

Hepatitis B virus (HBV) infection remains a major public health challenge globally, with approximately 254 million people living with chronic HBV, including over 6 million children under 5 years. Mother-to-child transmission (MTCT) of HBV is responsible for a significant portion of new infections, particularly in high-prevalence regions. Infants born to HBV-infected mothers are at risk of chronic infection, which can lead to severe liver diseases later in life. Truvada (TDF), a nucleoside reverse transcriptase inhibitor, is a recommended antiviral treatment for both HBV and HIV and has shown potential in reducing MTCT of HBV. However, the cost-effectiveness of TDF for preventing MTCT in resource-limited settings like Botswana is not well established. This study aims to evaluate the feasibility and cost effectiveness of three distinct strategies for screening and managing HBV among pregnant women in Botswana . This study will use a cohort of pregnant women in Botswana by assessing three groups as follows; i)No HBV screening or treatment is provided, and TDF prophylaxis is not administered (control group), ii)Screening for Hepatitis B surface antigen (HBsAg) is conducted for all pregnant women, with TDF prophylaxis administered to those who test positive for HBsAg, beginning at 28 weeks gestation and continuing for four weeks postpartum, iii)Screening for both HBsAg and HBV e-antigen (HbeAg) is performed, and TDF prophylaxis is administered exclusively to women who test positive for both HBsAg and HBeAg. A cost-utility analysis (CUA) will be conducted to compare the costs and clinical outcomes of each strategy, with effectiveness measured in terms of the number of HBV transmissions prevented. Costs will include screening for HBsAg and HBeAg, TDF treatment, hepatitis B immunoglobulin (HBIG) for infants, all components of the intervention (such as training, administration, supervision, etc) and maternal healthcare. In addition, a

decision-analytic model that would allow the generation of cost-effectiveness estimates will be designed. The Incremental Cost-Effectiveness Ratio (ICER) will be calculated to assess the cost per case of HBV transmission prevented for each strategy. Moreover, sensitivity analyses will be performed to test the robustness of results under varying assumptions related to drug costs, screening effectiveness, and intervention costs.

17: Application of Machine Learning Methods for the Analysis of Randomised Controlled Trials: A Systematic Review

Xiao Xuan Tan, Rachel Phillips, Mansour Taghavi Azar Sharabiani

Imperial College London

Background Randomised controlled trials (RCTs) collect extensive data on adverse events (AEs), yet their analysis and presentation are often overly simplistic, leading to missed opportunities for identifying potential signals of treatment-related harm. A 2024 scoping review identified a variety of machine learning (ML) approaches being employed in RCTs to identify heterogeneous treatment effects (HTEs) across key participant subgroups [1]. This highlights the range of ML methods being explored to derive insights from RCT data. ML methods hold potential to enhance AE analysis, offering tools to better interpret complex AE data and support data-driven, personalised treatment harm profiles. This review aims to identify ML methods and evaluate applications for analysis of RCT data, revealing both established and potentially suitable ML approaches that could be adapted to analyse AE data in RCTs. Additionally, this review will highlight emerging trends in ML applications to RCTs, including shifts in commonly used techniques, evolving best practices, and expanding use cases beyond HTE analysis.

Methods A systematic search was conducted in November 2024 via the Embase, MEDLINE, Web of Science and Scopus databases, alongside the preprint repositories arXiv, medRxiv and bioRxiv. Articles were eligible if they applied ML methods to analyse or reanalyse RCT datasets, irrespective of the types of outcomes examined, and accounted for the RCT's treatment assignment in their analyses. Following screening, a pre-piloted data extraction sheet will be used to systematically collect relevant study details.

Results After deduplication, 11286 articles were retrieved. Following title and abstract review, 2015 articles were eligible for full text review. Data extraction and synthesis are underway. Results presented will describe (i) study characteristics (e.g., purpose of analysis), (ii) RCT characteristics (e.g., medical area, trial design, outcomes examined) (iii) ML methods used, including model implementation details, use of explainability tools (e.g., SHAP,

LIME), results, limitations, reproducibility considerations (e.g., software, code availability, dataset access).

Conclusion The findings of this review will provide a comprehensive overview of applications of ML methods in RCTs, guiding trialists in their potential use for future trial design and analysis. Additionally, it will pinpoint ML techniques most relevant to the analysis of AEs, an area where more advanced analytical approaches are needed to facilitate early identification of potential signals of harm and improve the understanding of treatment-related harm.

[1] Inoue K, et al. Machine learning approaches to evaluate heterogeneous treatment effects in randomized controlled trials: a scoping review. *Journal of Clinical Epidemiology*. 2024; 176: 111538.

18: Joint Longitudinal Modelling of Non-Normally Distributed Outcomes and Endogenous Covariates

Chiara Degan¹, Bart Mertens¹, Pietro Spitali², Erik H. Niks², Jelle Goeman¹, Roula Tsonaka¹

¹Department of Biomedical Data Sciences, Leiden University Medical Center, The Netherlands,

²Department of Human Genetics, Leiden University Medical Center, The Netherlands,

In biomedical research, longitudinal outcomes and endogenous time-dependent covariates are often recorded, creating the need to develop methodological approaches to assess their associations, evaluate how one outcome changes in relation to the covariate, and determine how this relationship evolve over time.

To address these aspects, endogenous covariate and outcome are typically modelled jointly by assuming correlated random effects (Verbeke et al., 2014). We refer to this model as the Joint Mixed Model (JMM). This approach allows for the combination of variables of different types, while preserving the parameter interpretation of the univariate case and accommodating unbalanced data. However, the association is interpreted through the correlation of random effects rather than directly on the scale of the observed variables. Moreover, by making assumptions on the form of the variance-covariance matrix of the random effects, we impose some constraints on the variables' association form that may lead to biased estimations if misspecified.

As an alternative, we consider a modification of the joint model proposed by Rizopoulos

(2017), adapting it to include only longitudinal outcomes rather than a time-to-event component. We refer to this adapted model as the Joint Scaled Model (JSM). It induces the association by copying and scaling the linear predictor of the endogenous covariate into the linear predictor of the outcome. This approach preserves the advantages of the JMM while improving interpretability.

To compare the two model results and assess the impact of their underlying assumptions on conclusions, we propose to analytically derive an association coefficient that measures the marginal relation between variables. The purpose of this coefficient is to construct a quantity that has the same meaning in both models and can be interpreted similarly to a regression coefficient. It measures the change in outcome in response to a unit change in the covariate. Furthermore, it is a quantity that depends on the time of both variables, allowing it to capture the cross-sectional effect of the endogenous covariate on the outcome, as well as their relationship at different time points (lag effect).

The practical application of these models is limited by computational costs, which arise from high-dimensional integrations over random effects. To fill this gap, a flexible Bayesian estimation approach, known as INLA, has been used.

We will present the results of a longitudinal study on Duchenne Muscular Dystrophy patients, with a focus on evaluating the relationship between a bounded outcome and blood biomarkers.

19: Mediation Analysis for Exploring Gender Differences in Mortality among Acute Myocardial Infarction

Alice Bonomi, Arianna Galotta, Francesco Maria Mattio, Lorenzo Cangiano, Giancarlo Marenzi

IRCCS Centro Cardiologico Monzino, Italy

Background Women with acute myocardial infarction (AMI) have higher mortality rates than men, influenced by factors such as older age, comorbidities, atypical symptoms, and treatment delays. This study analyzed AMI patients (2003-2018) from the Lombardy Health Database (Italy) to investigate sex differences in in-hospital and one-year mortality, assessing the impact of age, percutaneous coronary intervention (PCI), and post-discharge therapy using mediation analysis.

Methods Among 263,564 AMI patients (93,363 women, 170,201 men), the primary and

secondary endpoints were in-hospital and one-year mortality, respectively. Mediation analysis was performed to evaluate the direct and indirect effects of sex on outcomes, incorporating age, PCI, and post-discharge therapy as mediators. The analysis was conducted using the SAS Proc CALIS procedure (SAS Institute Inc., Cary, NC, USA) based on structural equation modeling, with relationships quantified using standardized β coefficients.

Results Women had significantly higher in-hospital mortality (10% vs. 5%; $P<0.0001$) and one-year mortality (24% vs. 14%; $P<0.0001$) compared to men. Mediation analysis revealed that female sex directly contributed 12% to in-hospital mortality and 4% to one-year mortality, whereas age and undertreatment accounted for the majority of the disparity (88% [$\beta=0.09$] and 96% [$\beta=0.15$], respectively).

Conclusion Women with AMI experience higher mortality, primarily due to older age and undertreatment, both during hospitalization and after discharge. Addressing these disparities through optimized treatment strategies may improve outcomes in women with AMI.

20: Bivariate Random-Effects Models for the Meta-Analysis of Rare Events

Danyu Li, Patrick Taffe

Center for Primary Care and Public Health (unisanté), Division of Biostatistics, University of Lausanne (UNIL), Switzerland

It is well known that standard methods of meta-analysis, such as the inverse variance or DerSimonian and Laird methods, break down with rare binary events. Not only are effect sizes and within-study variances badly estimated, but also heterogeneity is generally not identifiable or strongly underestimated, and the overall summary index is biased. Many alternative estimation methods have been proposed to improve the estimates in sparse data meta-analysis. In addition to the Bivariate Generalized Linear Mixed Model (BGLMM), the Marginal Beta-Binomial, and the Sarmanov Beta-Binomial models are competitive alternatives. These models have already been used in the context of meta-analysis of diagnostic accuracy studies, where the correlation between sensitivity and specificity is likely to be strongly negative. To our best knowledge, they have not been investigated in the context of rare events and sparse data meta-analysis with a focus on estimating the Risk Difference (RD), Relative Risk (RR), and Odds Ratio (OR). Therefore, the goal of this study was to assess the performance and robustness of these three competitive models in this context. More specifically, the robustness of each model will be assessed using data-generating processes based on the other two competing models. For example, if the data were simulated based on the Sarmanov distribution, then the BGLMM and Marginal Beta Binomial models

are misspecified, and assessing their robustness is of interest. According to the simulation results, the BGLMM performs worst regardless of the misspecification of the distribution. The Sarmanov Beta-Binomial model and the Marginal Beta-Binomial model perform better and are more stable due to their lower variance.

21: Time-Varying Decomposition of Direct and Indirect Effects with Multiple Longitudinal Mediators

Yasuyuki Okuda¹, Masataka Taguri²

¹Daiichi Sankyo Do., Ltd., Japan

²Tokyo Medical University

Recent advances in mediation analysis using causal inference techniques have led to the development of sophisticated methods for complex scenarios, including those involving multiple time-varying mediators. Although these approaches accommodate time-varying mediators, their estimates are typically restricted to a single timepoint of interest, thus limiting our understanding of the temporal dynamics of mediation processes. In many clinical contexts, it is essential to capture how mediator effects vary over time to elucidate underlying mechanisms and optimize intervention timing. For example, temporal variations in direct and indirect effects can reveal critical windows during which a treatment exerts its primary influence. To address these limitations, we proposed a novel framework that extends existing approaches based on interventional direct and indirect effects with multiple time-varying mediators and treatment-mediator interaction.

Our method not only decomposes the overall effect into direct and indirect effects, but also further decomposes these effects into time-varying components to investigate mediated effects both up to and beyond the timepoint (t), thereby capturing their longitudinal trajectories. We also proposed a practical estimation approach using marginal structural models (MSMs) for both the outcome and mediators, using inverse probability weighting (IPW) method to account for time-varying confounders.

To illustrate the utility of our method, we applied it to the data from a randomized controlled trial evaluating the effect of a mineralocorticoid receptor (MR) blocker on urinary albumin-to-creatinine ratio (UACR) reduction. Specifically, we investigated how much of the treatment effect is mediated by changes in blood pressure and renal function (measured by eGFR) and explored differences in their mediator-specific effects over time. Our analysis indicated that the mediated effects via both systolic blood pressure and eGFR were relatively small compared with other pathways, with different patterns observed in their longitudinal trajectories.

We believe our approach provides investigators with a valuable tool for understanding an agent's mechanism of action, distinguishing it from other agents, and ultimately informing treatment decisions appropriate for each patient.

22: Causal Framework for Analyzing Mediation Effects of Clinical Biomarkers

Jinesh Shah

CSL Behring, Germany

For a biomarker to be at least a "level 3 surrogate" that is "reasonably likely to predict clinical benefit for a specific disease and class of interventions" [1] it must be either a mediator [1,2] on the causal pathway between treatment and response, or else be causally downstream of such a mediator. We investigate causal mediation analysis as an approach to statistically infer potential mediation effects of biomarkers. Steps involve graphically stating the causal structure using DAGs, formulating estimands of interest and using statistical methods to derive estimates. However, longitudinal clinical data are commonplace and causal estimation of such data is notoriously challenging, standard statistical methods might not provide appropriate target estimates. Thus, we also explore methods to account for time-varying confounding in mediation analysis, one such method discussed provides a reasonable approximation by "Landmarking" the biomarker process at a particular timepoint t [3], and modeling the clinical outcome data after time t . We aim to outline fundamental ideas of causal mediation [4] analysis and delineate a potential framework for its use in clinical development.

- (1) Fleming, T.R. and Powers, J.H. Biomarkers and surrogate endpoints in clinical trials. *Stat. Med.* 31 (2012):2973–2984.
- (2) Joffe, M.M. and Greene, T. Related causal frameworks for surrogate outcomes. *Biometrics* 65 (2009):530–538.
- (3) Putter, H. and van Houwelingen, H.C. Understanding landmarking and its relation with time-dependent Cox regression. *Stat. Biosci.* 9 (2017):489–503.
- (4) Imai, K., Keele, L. and Tingley, D. A general approach to causal mediation analysis. *Psychol. Methods* 15 (2010):309–334.

23: Retracted.

24: DoubleMLDeep: Estimation of Causal Effects with Multimodal Data

Martin Spindler^{1,3}, Victor Chernozhukov², Philipp Bach¹, Jan Teichert-Kluge¹, Sven Klaassen^{1,3}, Suhas Vijaykumar²

¹Universität Hamburg, Germany

²MIT, USA

³Economic AI, Germany

This paper explores the use of unstructured, multimodal data, namely text and images, in causal inference and treatment effect estimation. We propose a neural network architecture that is adapted to the double machine learning (DML) framework, specifically the partially linear model. An additional contribution of our paper is a new method to generate a semi-synthetic dataset which can be used to evaluate the performance of causal effect estimation in the presence of text and images as confounders. The proposed methods and architectures are evaluated on the semi-synthetic dataset and compared to standard approaches, highlighting the potential benefit of using text and images directly in causal studies. Our findings have implications for researchers and practitioners in medicine, biostatistics and data science in general who are interested in estimating causal quantities using non-traditional data.

25: Causal Machine Learning Methods for Estimating Personalised Treatment Effects - Insights on Validity from Two Large Trials

Hongruyu Chen, Helena Aebersold, Milo Alan Puhan, Miquel Serra-Burriel

University of Zurich, Switzerland

Causal machine learning (ML) methods hold great promise for advancing precision medicine by estimating personalised treatment effects. However, their reliability remains largely unvalidated in empirical settings. In this study, we assessed the internal and external validity of 17 mainstream causal heterogeneity ML methods—including metalearners, tree-based methods, and deep learning methods—using data from two large randomized controlled trials: the

International Stroke Trial (N=19,435) and the Chinese Acute Stroke Trial (N=21,106). Our findings reveal that none of the ML methods reliably validated their performance, neither internal or external, showing significant discrepancies between training and test data on the proposed evaluation metrics. The individualized treatment effects estimated from training data failed to generalize to the test data, even in the absence of distribution shifts. These results raise concerns about the current applicability of causal ML models in precision medicine, and highlight the need for more robust validation techniques to ensure generalizability.

26: Challenges with Subgroup Analyses in Individual Participant Data Meta-Analysis of Randomised Trials

Alain Amstutz^{1,2,3}, Dominique Costagliola⁴, Corina S. Rueegg^{2,5,6}, Erica Ponzi^{2,5}, Johannes M. Schwenke¹, France Mentré^{7,8}, Clément R. Massonnaud^{7,8}, Cédric Laouénan^{7,8}, Aliou Baldé⁴, Lambert Assoumou⁴, Inge C. Olsen^{2,5}, Matthias Briel^{1,9}, Stefan Schodelmaier^{9,10,11}

¹Division of Clinical Epidemiology, Department of Clinical Research, University Hospital Basel and University of Basel, Basel, Switzerland

²Oslo Centre for Biostatistics and Epidemiology, Oslo University Hospital, Oslo, Norway

³Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom

⁴Sorbonne Université, Inserm, Institut Pierre-Louis d'Épidémiologie et de Santé Publique, Paris, France

⁵Department of Research Support for Clinical Trials, Oslo University Hospital, Oslo, Norway

⁶Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Zurich, Switzerland

⁷Université Paris Cité, Inserm, IAME, Paris, France

⁸Département d'Épidémiologie, Biostatistique et Recherche Clinique, Hôpital Bichat, AP-HP, Paris, France

⁹Department of Health Research Methods, Evidence, and Impact (HEI), McMaster University, Hamilton, Canada

¹⁰School of Public Health, University College Cork, Cork, Ireland

¹¹MTA–PTE Lendület "Momentum" Evidence in Medicine Research Group, Medical School, University of Pécs, Pécs, Hungary

Background Individual participant data meta-analyses (IPDMA) offer the opportunity to conduct credible subgroup analyses of randomized clinical trial data by standardising subgroup definitions across trials, avoiding between-trial information sharing, and enabling effect comparison from trial to trial. These advantages are reflected and judged in item 1 and 2 of

the Instrument for the Credibility of Effect Modification ANalyses (ICEMAN), a tool increasingly used by Cochrane meta-analysts and the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group. However, guidance on optimal approaches to inform these ICEMAN items when conducting an IPDMA is limited and might differ when using an IPDMA one-stage versus two-stage models. We recently conducted two large IPDMAs, analysing 20 COVID-19 trials with over 23,000 randomised participants. Here, we provide a case report on the approaches used to inform ICEMAN item 1 and 2.

Methods Following a pre-specified protocol, we applied one- and two-stage models for these IPDMAs, and documented challenges and mitigation strategies along the subgroup analysis process to enhance guidance for future updates to the ICEMAN tool.

Results We identified several challenges. First, ensuring that the one-stage model separates within from between trial information (ICEMAN item 1), as the two-stage model does by design, is difficult and requires stratification of certain parameters, and correctly specifying the random parameters. Second, the default estimation methods may differ depending on the statistical packages used for one- and two-stage, resulting in different interaction estimates to inform ICEMAN item 1. Third, choosing descriptive thresholds for continuous effect modifiers in meta-analysis of interaction plots can mislead about the direction of effect modification in individual trials (ICEMAN item 2). We developed illustrative modular R codes to inform ICEMAN item 1 with one- and two-stage models, and provided plots with meta-analysis of interaction estimates alongside trial-specific subgroup effects to inform ICEMAN item 2.

Conclusion At the conference, we will present these challenges in detail, their mitigation strategies and discuss the need for refining methods guidance to evaluate the effect modification credibility in IPDMAs using the ICEMAN tool.

27: Illustration and Evaluation of a Causal Approach to Sensitivity Analysis for Unmeasured Confounding using Measured Proxies with a Simulation Study

Nerissa Nance^{1,2}, Romain Neugebauer³

¹Novo Nordisk, Denmark

²University of California, Berkeley CA

³Kaiser Permanente Northern California Division of Research, Pleasanton CA

Introduction Sensitivity analysis for unmeasured confounding is a key component of applied causal analyses using observational data[1]. A general method [2] based on a rigorous causal framework has been previously proposed; this approach addresses limitations of existing meth-

ods such as reliance on arbitrary parametric assumptions or expert opinion without taking advantage of the available data at hand. We illustrate and evaluate this general method through a simulation study.

Methods

We simulated data using a parametrized nonparametric structural equation model. Our simulated observed data consisted of unmeasured covariate, measured covariate, exposure, and outcome. We studied the performance of point and interval estimation of an inverse probability weighting estimator that aims to adjust for unmeasured confounding through a measured proxy variable. We assessed this method under a range of scenarios, including: interaction terms with the exposure, various association strengths and directions between the covariates and the exposure/outcome.

Results We demonstrated potential bias elimination and recovery of confidence interval coverage from unmeasured confounding in the case where the unmeasured covariate has the same magnitude and direction of association with both exposure and outcome as the measured proxy. However, in other scenarios, such as when the measured and unmeasured confounders had antagonistic effects, recovery was low or minimal.

Discussion We illustrate through simulations that when there is the same magnitude and direction of the association of the unmeasured confounder and measured proxy with the exposure and outcome, the true unconfounded effect can be fully recovered. However, we also show how this recovery can break down in other situations that analysts may encounter. Results from this study informs key practical considerations for applying these methods, as well as highlight potential limitations.

References

1. Dang LE et al.. A causal roadmap for generating high-quality real-world evidence. *J Clin Transl Sci.* 2023 Sep 22;7(1):e212.
2. Luedtke, A.R., Diaz, I. and van der Laan, M.J., 2015. The statistics of sensitivity analyses.

28: Quantifying Causal Treatment Effect on Binary Outcome in RCTs with Noncompliance: Estimating Risk Difference, Risk Ratio and Odds Ratio

Junxian Zhu, Mark Y. Chan, Bee-Choo Tai

National University of Singapore, Singapore

Randomized Controlled Trials (RCT) are currently the most reliable method for empirically evaluating the effectiveness of a new drug. However, patients may fail to adhere to the treatment protocol due to side effects. Medical guidelines recommend reporting the risk difference (RD), the risk ratio (RR) and the odds ratio (OR), as they offer distinct perspectives on the effect of the same drug. Unlike RD, there are only a few available methods to estimate RR and OR for RCT in the presence of non-compliance. In this paper, we propose a new inverse probability weighting (IPW)-based RD, RR and OR estimators for RCT in the presence of non-compliance. This IPW-based method creates a new categorical variable by utilizing information on non-compliance with the randomly assigned treatment. For all estimators, we prove their identification, asymptotic normality and derive corresponding asymptotic confidence intervals. We evaluate the performance of these three estimators through an intensive simulation study. Its application is further demonstrated using data from the IMMACULATE trial on remote post-discharge treatment for patients with acute myocardial infarction.

29: Blinded Sample Size Recalculation for Randomized Controlled Trials with Analysis of Covariance

Takumi Kanata, Yasuhiro Hagiwara, Koji Oba

Department of Biostatistics, School of Public Health, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

Background / Introduction In randomized controlled trials, covariate adjustment can improve the statistical power and reduce the necessary sample size compared to unadjusted estimator. Analysis of covariance (ANCOVA) is often used for adjusting baseline covariates when outcome is continuous. For designing a sample size based on ANCOVA, it is necessary to pre-specify the association between outcome and baseline covariates, as well as that among baseline covariates. However, determining these parameters at the design stage is challenging. While it may be possible to adaptively assess them during the trial, the statistical impact remains unclear. In this study, we propose a blinded sample size recalculation method for ANCOVA estimator, which is asymptotically valid under minimum distributional assumptions and thus allows for arbitrary model misspecification.

Methods We show that the asymptotic variance of ANCOVA estimator and unadjusted estimator can be calculated using the pooled outcome and baseline covariates when the treatment is randomly assigned with 1:1 ratio independent of the baseline covariates. This result is valid under arbitrary model misspecification. Our proposal is as follows. First, we

calculate the sample size based on a t-test without adjusting for baseline covariates. Then, at a specific time point (e.g. when 50% of outcome is observed), we assess the relevant parameters under blinded conditions without examining the between-group differences. We propose a sample size recalculation method that considers the asymptotic variance reduction through covariate adjustment and recalculate the final sample size based on this proposed method. We conducted simulations to evaluate the performance of the proposed method under various scenarios.

Results The proposed method achieved a nominal statistical power under various scenarios and it reduced the necessary sample size at the final analysis according to the correlations between the outcome and the baseline covariates; for example, when the correlations are 0.5, the sample size reduction ranged from 15% to 36% on average. Although the proposed method was based on the asymptotic results, it performed well under the relatively small sample size. We also found that type-I error at the final analysis was not affected by the proposed method.

Conclusion The proposed sample size recalculation method achieves a nominal statistical power in randomized controlled trials based on ANCOVA without type-I error inflating. The proposed method possibly reduces the necessary sample size, and it would lead to efficient drug development.

30: Variance Stabilization Transformation for the Intraclass Correlation Coefficient of Agreement with an Application Example to Meta-Analyses of Inter-Rater Reliability Studies

Abderrahmane Bourredjem^{1,2,3}, Isabelle Fournel¹, Sophie Vanbelle⁴, Nadjia El Saadi³

¹Inserm CIC1432, Centre d'investigation clinique, Module Epidémiologie Clinique/Essais cliniques, CHU de Dijon, France.

²Institut de Mathématiques de Bourgogne, UMR 5584, CNRS, Université de Bourgogne, F-21000 Dijon, France.

³LAMOPS, École Nationale Supérieure de Statistique et d'Economie Appliquée, Kolea, Algérie.

⁴Department of Methodology and Statistics, Care and Public Health Research Institute (CAPHRI), Maastricht University, The Netherland.

Introduction :

We consider the problem of variance stabilizing transformation (VST) for the two-way intra-

class correlation coefficient of agreement (ICC2a) in inter-rater reliability studies when both raters and subjects are assumed to be randomly selected from their respective populations. Such transformations aim to make the ICC2a's variance independent from its estimate, improving the ICC2a confidence interval (CI) and the combination of independent ICC2as in meta-analyses. In this work, we calculate three potential VSTs for the ICC2a, evaluate their properties by simulation for single study CIs and demonstrate their use on meta-analysis of inter-rater reliability studies.

Methods :

It was recently shown that the variance of the ICC2a estimate depends on a nuisance parameter, defined as the ratio of the inter-rater to the inter-subject variances. Using this variance expression, three VST approximations (noted T0, T1 and T2) were obtained addressing the nuisance parameter differently. A simulation study with small to moderate sample sizes compared the properties of the obtained VSTs against two reference CIs methods: 1) the modified large sample approach (MLS), 2) a beta-distribution-based method (β). Finally, we illustrated the use of our VSTs on a single inter-rater reliability study with 10 physiotherapists evaluating the exercise performance of 42 low back pain patients, as well as on a meta-analysis of 11 inter-rater reliability studies of upper extremity muscle tone measurements.

Results :

The analytical expression of the three VSTs varies in complexity, from T0 (the simplest) to T2 (the most complex, requiring numerical methods to calculate its inverse transformation), through T1 (a middle level between T0 and T2). Simulations show that for small samples (up to 30 subjects and fewer than 10 raters), the MLS and β approaches remain preferable. For medium-sized samples (from 40 subjects and 10 raters), T1 provides coverage rates close to 95% while shortening the CI length. In the meta-analysis example, T1 offers advantages including transformed estimators simpler to interpret and better considering study weights in the synthesis of the ICC2a estimates and their CI.

Conclusion :

We propose a novel VST (noted T1) for ICC2a, filling a gap in the literature. We recommend using T1 for ICC2a CIs in medium-sized individual studies and for meta-analyses of inter-rater reliability studies. However, more extensive simulations are required to refine this recommendation, especially for meta-analyses.

31: Bridging Single Arm Studies with Individual Participant Data in Network Meta-Analysis of Randomized Controlled Trials: A Simulation Study

Katerina Maria Kontouli, Stavros Nikolakopoulos, Christos Christogiannis, Dimitrios Mavridis

University of Ioannina, Greece

Background There is a growing interest in including single-arm studies within health technology assessments (HTA). Manufacturers often have access to individual participant data (IPD) from their own studies (a single-arm study evaluating treatment B), while only aggregate data (AGD) are available from published studies (e.g., comparing treatments C, D etc. to a reference treatment A). Several methods such as the Matching-Adjusted Indirect Comparison (MAIC) and the Simulated Treatment Comparison (STC) have been suggested to estimate an indirect effect (e.g, BvsA, BvsC) when the distribution of prognostic factors and effect modifiers differ across studies. The aim is to evaluate MAIC and STC in estimating an indirect effect in the above scenario through a simulation study.

Methods We examined three methods: two widely used adjusted methods for unanchored comparisons, MAIC and STC, and the naïve (unadjusted) method. We applied these methods to incorporate single-arm studies with available interventions within a connected network of randomized controlled trials. To optimize the matching process, we employed two distinct distance metrics: Gower's and Mahalanobis distance. Our simulation study explored various scenarios, varying (i) the sample size of studies, (ii) the magnitude of the treatment effect, (iii) the correlation between continuous covariates representing study population characteristics, (iv) the baseline probability, and (v) the degree of overlap between the single-arm study and the RCTs.

Results Our simulation results indicate that when all continuous covariates are drawn from the same distributions with zero correlation, all methods perform similarly in terms of bias, mean squared error, and coverage across all scenarios. However, when the covariate overlaps between the single-arm study and the RCTs is around 80%, the Bucher method produces more biased estimates compared to MAIC and STC. As the overlap decreases to approximately 60%, the differences between MAIC and STC become more pronounced, particularly in terms of coverage and MSE.

Conclusion STC emerges as the most robust approach for integrating evidence from single-arm studies into a network of RCTs. Additionally, Mahalanobis distance proves to be effective in identifying the optimal match, enhancing the reliability of the synthesis.

32: Comparative Efficacy and Safety of Migraine Treatments: A Network Meta-Analysis of Clinical Outcomes

Shashank Tripathi¹, Rachna Agarwal²

¹University College of Medical Sciences GTB Hospital, New Delhi, India

²Institute of Human Behavior and Allied Sciences, New Delhi, India

Introduction Migraine is a common and debilitating neurological condition, affecting roughly 10% of the global population and placing a significant burden on public health. It occurs in episodes, often characterized by intense headaches accompanied by sensitivity to light (photophobia), sensitivity to sound (phonophobia), and a range of autonomic and sensory disturbances.

Methods A comprehensive search of three databases was conducted up to April 30, 2023. A frequentist network meta-analysis was utilized to estimate both direct and indirect effects across three outcomes; mean migraine days, freedom for pain in two hours, and adverse event. Interventions were ranked independently for each outcome using the p-score. The choice of meta-analysis model was based on the I^2 statistic: a random-effects model was applied when I^2 exceeded 30%, while a fixed-effect model was used when I^2 was 30%. All statistical analyses were performed using R version 4.3.2.

Results A total of 80 articles were included in current investigation. For, change in mean migraine days (MMD) as direct estimate, suggested statistically significant result for CGRP antagonist [SMD: -0.38 (-0.61, -0.14)], CGRP mAbds [SMD: -0.35 (-0.41, -0.31)] and Triptans [SMD: -0.36 (-0.62, -0.10)]. Similarly, direct estimates were calculated for freedom for pain in two hours, suggested statistically significant result CGRP antagonist [RR: 5.83 (2.50, 13.59)], Dihydroergotamine [RR: 19.92 (3.41, 116.76)], Nasal agent (NSAID) [RR: 10.27 (1.03, 102.28)], Nasal agent (Triptan) [RR: 8.27 (3.51, 19.56)], NSAID [RR: 19.1 (7.36, 49.01)], and Triptan [RR: 22.82 (16.74, 31.12)]. Additionally, for the outcome adverse event, the direct estimate suggested statistically significant result for CGRP mAbs [RR: 2.77 (1.97, 3.91)], CGRP antagonist [RR: 2.92 (1.95, 4.37)], Dihydroergotamine [RR: 3.99 (1.47, 10.82)], NSAID [RR: 4.21 (2.1, 8.1)], Nasal agent (CGRP antagonist) [RR: 7.61 (2.31, 25.19)], Triptans [RR: 8.40 (6.91, 10.22)], Nasal agent (triptan) [RR: 23.57 (9.01, 61.71)]. The indirect estimates were calculated taking all treatments under investigations as reference treatment, simultaneously, for each outcome of interest.

Conclusion A network meta-analysis of migraine treatments found Triptans to be highly effective for pain, though with a higher risk of adverse events. CGRP antagonists excelled at reducing monthly migraine days but also had increased side effects.

33: Optimal Standardization as an Alternative to Matching using Propensity Scores

Ekkehard Glimm, Lillian Yau

Novartis Pharma, Switzerland

In many development programs in the pharmaceutical industry, there is a need for indirect comparisons of medical treatments that were investigated in separate trials. Usually, trials have slightly different inclusion criteria, hence the influence of confounding factors has to be removed for a “fair” comparison. The most common method applied for this is propensity score matching. This method yields a set of weights used to re-weight patients in such a way that the weighted averages of the confounding variables are rendered comparable across the studies.

Propensity score matching typically achieves “roughly matched” groups, but almost invariably some differences between the averages of the matching variables in the compared trials remain.

We have recently suggested an approach for exact matching which may serve as an alternative to propensity score matching via a logistic regression model. This approach treats the matching problem as a constrained optimization problem. This approach guarantees that post-matching, the averages of the variables used in matching from the two trials are identical. While several objective functions could in theory be selected to generate a set of weights, in this talk we will focus on weights that maximize the effective sample size (ESS).

While the approach is closely related to matching-adjusted indirect comparison (MAIC, Signorovitch et al, 2010), it goes beyond their suggestion because we do not impose a specific functional form on the matching weights. Furthermore, in the talk we focus on the case where individual patient data (IPD) is available from all trials in the analysis, whereas the original MAIC approach considered only the matching of IPD onto aggregated data.

In the talk, we illustrate the application of the approach to two studies. Furthermore, we present the results from a simulation study showing that the new suggestion leads to weights which are considerably more stable than propensity score weights.

References Signorovitch JE, Wu EQ, Andrew P, et al. Comparative effectiveness without head-to-head trials: a method for matching-adjusted indirect comparisons applied to psoriasis treatment with adalimumab or etanercept. *PharmacoEconomics*. 2010;28(10):935-945.

Glimm, E. and Yau, L. (2022): Geometric approaches to assessing the numerical feasibility

for conducting matching-adjusted indirect comparisons. *Pharm Stat* 21, 974-987.

34: Evaluating Diagnostic Tests Against Composite Reference Standards: Quantifying and Adjusting for Bias

Vera Hudak, Nicky J. Welton, Efthymia Derezea, Hayley Jones

University of Bristol, United Kingdom

Background Composite reference standards (CRSs) are often used in diagnostic accuracy studies in situations where gold standards are unavailable or impractical to carry out on everyone. Here, the test under evaluation is compared with some combination (composite) of results from other tests. We consider a special case of CRS, which we refer to as a 'check the negatives' design. Here, all study participants receive an imperfect reference standard, and those who test negative on this are additionally tested with the gold standard. Unless the imperfect reference standard is 100\% specific, some bias can be anticipated.

Methods We derive algebraic expressions for the bias in the estimated accuracy of the test under evaluation in a 'check the negatives' study, under the assumption that test errors are independent given the true disease status. We then describe how bias can be adjusted for using a Bayesian model with an informative prior for the specificity of the imperfect reference standard, based on external information. Our approach is evaluated through a simulation study under two scenarios. First, we consider the case where the prior for the specificity of the imperfect reference standard is correctly centred around its true value, and we assess the impact of increasing uncertainty by increasing the prior standard deviation. Second, we examine the case where the prior is incorrectly centred, but the true value remains within the 95\% prior credible interval, to explore the consequences of moderate prior misspecification.

Results/Conclusions: In a 'check the negatives' study, under the assumption of conditional independence of errors made by the test under evaluation and the imperfect reference standard, the estimated specificity is unbiased but the sensitivity is underestimated. Preliminary findings suggest that, if the informative prior is correctly centred, the Bayesian model will always reduce bias and can successfully eliminate it in some, but not all, scenarios. Full simulation results, including those with incorrectly centred prior, and their implications will be presented at the conference.

35: Characteristics, Design and Statistical Methods in Platform Trials: A Systematic Review

Clément R. Massonnaud^{1,2}, Christof Manuel Schönenberger³, Malena Chiaborelli³, Selina Ehrenzeller³, Alexandra Griessbach³, André Gillibert⁴, Matthias Briel³, Cédric Laouénan^{1,2}

¹Université Paris Cité, Inserm, IAME, F-75018 Paris, France

²AP-HP, Hôpital Bichat, Département d'Épidémiologie, Biostatistique et Recherche Clinique, F-75018 Paris, France

³CLEAR Methods Center, Division of Clinical Epidemiology, Department of Clinical Research, University Hospital Basel, University of Basel, Basel, Switzerland

⁴Department of Biostatistics, CHU Rouen, Rouen, France

Background

Platform trials (PTs) are gaining popularity in clinical research due to their innovative and flexible methodologies. However, their complex design underscores the need for a review of how they are currently implemented. The objective of this systematic review was to determine the characteristics, methodological and statistical practices in PTs.

Methods We identified PTs from trial registries and bibliographic databases up to August 2024. Eligible PTs were randomized controlled trials studying multiple interventions within a single population, with flexibility to add or drop arms. Data were extracted on trial status, design, statistical methods, and reporting practices. Key variables included sample size determination, interim analyses, and type I error control. Descriptive statistics summarized findings across therapeutic areas and statistical framework (frequentist or Bayesian).

Results We identified 190 PTs. Most focused on infectious diseases (77 [40.5%], including 57 for COVID-19) and oncology (69 [36.3%]). PT initiation peaked during the COVID-19 pandemic but has since stabilized at 85 active trials, with 25 PTs in planning. Non-industry sponsorship accounted for 78% (142/183) of PTs, with differences between infectious disease (95%, 71/75) and oncology trials (51%, 35/68). A complete master protocol was available for 47% (89/190) of all PTs and for 55% (83/152) of ongoing, completed, or discontinued PTs. Amendments were tracked in 61% (52/85) of protocols with multiple versions. Registry entries were considered up-to-date for 87% (153/175) of registered PTs. Bayesian designs featured in 59/190 PTs versus 56/190 frequentist trials, 20/190 trials utilizing both frameworks (unclear statistical framework in 55/190 PTs). Overall, 25/111 trials (23%) were designed without a pre-determined target sample size, all of which were Bayesian. Among these, 15 were explicitly reported as “perpetual” trials. The number of interim analyses was

pre-determined in 19% (11/58) of Bayesian trials versus 58% (28/48) of frequentist trials. Simulations to evaluate operating characteristics were used in 93% (39/42) of Bayesian trials. Simulation reports were available in 67% (26/39) of cases, and the procedure was detailed for 62% (24/39) of trials. Only two trials shared the simulation code.

Conclusions

Platform trials remain popular and increasingly diverse. Efforts to enhance transparency and reporting, especially in complex Bayesian platform trials, are essential to ensure reliability and broader acceptance.

36: WRestimates: An R Package for Win-Ratio Sample Size and Power Calculations

Autumn O Donnell

University of Galway, Ireland

The win-ratio has excellent potential for determining the overall efficacy of treatments and therapies in clinical trials. 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 outside of cardiovascular disease, there is a need for a straightforward approach to the study design, particularly the power and sample size determination. An appropriate method for determining these metrics is vital to ensure the validity of the results obtained in a study. The WRestimates package provides easy-to-use functions which can be used in required sample size determination and power of studies implementing the win-ratio. These allow for the calculation of sample size and power based on estimands or pilot data, negating the need for complex simulation-based methods which require many assumptions to be made of the data.

37: Randomizing with Investigator Choice of Treatment: A Powerful Pragmatic Tool in Clinical Trials

Lillian Yau, Betty Molloy

Novartis Pharma, Switzerland

Taking a patient-centric approach, pragmatic clinical trials aim for study designs that are closer to clinical practice. Results of treatment benefits and risks of new medical products from these trials can provide information for patients, health care professionals, and decision-makers that are more easily generalized to the real world.

We present as an example the design of a multi-regional, phase III registration study for a first-line cancer treatment. The study compares an experimental treatment against two generations of standards-of-care (SoC) that are approved and used worldwide for newly diagnosed patients. The first generation (1G) and second generation (2G) treatments differ with respect to efficacy and safety.

To mimic clinical practice, before randomization, trial investigators and patients together selected an SoC option based on patient-related factors such as age, comorbidities, disease characteristics, as well as on regional practice. This choice of SoC was used as a stratification factor in the randomization and subsequent data analysis. This approach facilitates causal inference on the comparison of the experimental treatment with the different SoC options (1G or 2G) separately and combined.

To satisfy the requirements of different health authorities and the reimbursement agencies, joint primary endpoints as well as key secondary endpoints were designed to be tested at different time points. Strong-control of the type I error was guaranteed by combining multiplicity adjustment with group sequential testing.

By allowing investigator choice of 4 currently available SoC in the active control arm, the study optimized patients' treatment and reduced the risk of exclusion of patients. It was very attractive to both patients and physicians, as reflected in the fast recruitment with close to 30 patients per month, nearly double what was expected in this disease area.

The study had its primary and key secondary read-outs in 2024. The primary results were the basis of the approval of the new treatment in many countries including the US, Canada, and Switzerland. The key secondary results are used to support the submission to EMA.

This study not only advances the therapeutic landscape but also sets a benchmark for future clinical trials, demonstrating that patient-centered strategies and robust designs can address the requirements of multiple decision makers and can lead to significant advancements in clinical research and patient care.

38: Confirming Assay Sensitivity in 2-Arm Non-Inferiority Trial using Meta-Analytic-Predictive Approach

Satomi Okamura¹, Eisuke Hida²

¹Department Of Medical Innovation, The University of Osaka Hospital, Japan

²Graduate School of Medicine, The University of Osaka, Japan

Introduction and Objective: Assay sensitivity is a well-known issue in 2-arm non-inferiority (NI) trials. To assess assay sensitivity, a 3-arm NI trial including placebo, control, and treatment is strongly recommended, with concerns about ethics and feasibility. FDA guidance on NI trials states: "In the absence of a placebo arm, knowing whether the trial had assay sensitivity relies heavily on external information (not within-study), giving NI studies some of the characteristics of a historically controlled trial." Hence, the new NI trial must be similar to the historical trials. Additionally, the historical trials must have consistently shown that the 'control' in the NI trial is superior to placebo. The superiority here requires that the effect of the 'control' minus a NI margin is greater than that of placebo, not just the 'control'.

Our objective is to propose a method to ensure the similarity of the NI trial to the historical trials and the superiority of 'control' to placebo for confirming assay sensitivity in the 2-arm NI trial. Information from historical trials usually consists of aggregate data. However, it has become clear that when effect modifiers are present, simple summary statistics for the entire population are insufficient. Therefore, it is important that the proposed method take into account the presence of effect modifiers.

Method and Results: To assess assay sensitivity, we use the meta-analytic-predictive approach. This approach is the Bayesian method so the prior distribution, especially in this study for the between-trial heterogeneity, is crucial. We assume the parameter follows the half-normal distribution for deviation or inverse-gamma distribution for variance. The performance of the approach is evaluated from two perspectives. First, we assess the influence of the prior setting on assay sensitivity by varying the amount of prior information about between-trial heterogeneity. Second, we demonstrate what trials may reduce assay sensitivity by setting multiple conditions for the historical trials, such as the number of historical trials, sample size, effect size, and the property of effect modifiers. For each scenario, we compute the posterior distribution of the 'control' effect and assess the performance of the method through joint power and type I error rate.

Conclusions Our simulation study suggests that the meta-analytic-predictive approach is one of the useful methods to evaluate assay sensitivity in 2-arm NI trial. Especially, the consideration of uncertainty, which is unique to the Bayesian approach, is of great benefit where only the aggregate data in the historical trials are available.

39: Adding Baskets to an Ongoing Basket Trial with Information Borrowing: When Do you Benefit?

Libby Daniells¹, Pavel Mozgunov¹, Helen Barnett³, Alun Bedding⁴, Thomas Jaki^{1,2}

¹MRC Biostatistics Unit, Cambridge University, United Kingdom

²Faculty of Informatics and Data Science, University of Regensburg, Regensburg, Germany

³Department of Mathematics and Statistics, Lancaster University, Lancaster, United Kingdom

⁴Roche Products, Ltd, Welwyn Garden City, United Kingdom

Innovation in trial designs has led to the development of basket trials in which a single therapeutic treatment is tested in several patient populations, each of which forms a basket. This trial design allows for the testing of rare diseases or subgroups of patients. However, limited basket sample sizes can cause a lack of statistical power and precision of treatment effect estimates. This is tackled through the use of Bayesian information borrowing.

To provide flexibility to these studies, adaptive features are desirable as they allow for pre-specified modifications to an ongoing trial. In this talk we focus on the incorporation of (a) newly identified basket(s) part-way through a study. We propose and compare several approaches for adding new baskets to an ongoing basket trial under an information borrowing structure and highlight when it is beneficial to add a new basket to an ongoing trial as opposed to running a separate investigation for them. We also propose a novel calibration for the decision criteria in basket trials that is robust with respect to false decision making. Results display a substantial improvement in power for a new basket when information borrowing is utilized, however, this comes with potential inflation of error rates. This inflation is reduced under the novel calibration procedure.

40: Optimizing Adaptive Trial Design to Ensure Robustness Across Varying Treatment Effect Assumptions

Valeria Mazzanti¹, Dirk Klingbiel²

¹Cytel Inc.

²Bristol Myers Squibb

Background Strong adaptive clinical trial design relies on several key aspects: experience in a therapeutic area; expertise in statistical methodology; and appropriate technology to assess

design robustness. A recent study design assessment for a compound under development in Hematology highlighted each of these aspects in an interesting way. The study's primary endpoint was Progression-Free Survival (PFS), though there was also strong interest in monitoring observed events for Overall Survival (OS), adding to the design's complexity. Our aim in this assessment and optimization process was to shorten the expected average study duration, while still ensuring appropriate statistical power to detect a minimally clinically viable treatment effect for this product.

Methods The original design targeted 90% power using a 1-sided alpha of 0.025 and 1:1 randomization approach. The design included one interim analysis after 40% of PFS events, assessing futility only. In our simulation plan, we varied the number of interim analyses (1 or 2 interim looks) and explored the impact of a variety of interim timings and types of assessment (futility and/or efficacy) on the expected number of events. We also employed a multi-state model to simulate PFS and OS events for each patient so that we could report how many OS events would be observed at each analysis of PFS. These variations resulted in over 5,000 parameter combinations that were ranked and scored in-line with the stated strategic priority of overall reduction in study duration, using industry-standard advanced statistical software.

Results We managed to optimize our design such that the required sample size fell by 85 patients and was 13 months shorter in duration on average. The optimized design included two interim analyses, in which both efficacy and futility were assessed. The additional efficacy evaluations led to a high probability of early stopping without compromising on overall power of the study.

Conclusion The results of the exploration highlighted the value of adding an efficacy stopping boundary and a second, later interim analysis both leading to savings in average sample size and average study duration. Additional explorations may include assigning statistical significance to evaluate the treatment's impact on the OS endpoint as well, and assessing the probability of success of the trial by sampling the events generated in our simulation from prior distributions identified from historical studies.

41: N-of-1 Trials to Estimate Individual Effects of Music on Concentration

Thomas Gärtner¹, Fabian Stolp¹, Stefan Konigorski^{1,2}

¹Hasso Plattner Institute for Digital Engineering, Germany

²Hasso Plattner Institute for Digital Health at Mount Sinai, Icahn School of Medicine at Mount Sinai, USA

Focus and concentration are influenced by various environmental factors, with music shown to impact cognitive performance. However, recent research highlights the individualized nature of the effect of music, as responses vary largely based on genre and personal preference. Traditional population-level studies often obscure these differences, whereas N-of-1 trials can provide a personalized approach that may be particularly suited for examining how self-selected music genres affect concentration.

This study presents the design of a series of N-of-1 trials investigating the individual effects of music on cognitive processes as primary outcome, measured using a digitally adjusted Stroop test. In the study, participants will select one music genre, with or without lyrics, as their intervention, which will be compared to silence as a baseline. Each participant will be randomly assigned to a sequence of 3-minute music listening periods (intervention, A) and 3-minute silent periods (control, B) in a two-cycle crossover design (ABAB or BABA). To minimize carryover effects and concentration loss, a 1-minute break is scheduled between blocks. After each block, participants will complete a brief questionnaire to assess self-reported concentration and stress levels. Additionally, physiological proxies for stress and cognitive load, including heart rate, electroencephalography (EEG), and pupil dilation, will be recorded. Intervention effects will be estimated using a Bayesian linear mixed models, with a primary focus on individual-level analyses and secondary analyses at the population level.

This study will provide valuable insights into the personalized effects of music on concentration, helping individuals optimize their cognitive performance. At the population level, it will identify variations in concentration effects across different music genres, contributing to the broader understanding of music as a cognitive intervention.

42: Simulation Study Examining Impact of Study Design Factors on Variability Measures

Laura Quinn^{1,2}, Jon Deeks^{1,2}, Yemisi Takwoingi^{1,2}, Alice Sitch^{1,2}

¹Department of Applied Health Sciences, University of Birmingham

²National Institute for Health and Care Research (NIHR) Birmingham Biomedical Research Centre, Birmingham, UK

Introduction

Interobserver variability studies in diagnostic imaging are crucial for assessing the reliability of imaging test interpretations between different observers. The design and conduct of these studies are influenced by various factors that can impact the calculation and interpretation

Abstracts of Contributed Posters

of variability estimates. These factors include participant sample size, condition prevalence, diagnostic test discrimination, and reader error levels.

Methods

Data was simulated for a study design with binary outcomes and two interpretations for each patient. A range of scenarios were simulated, varying participant sample size (25 to 200), condition prevalence (5% to 95%), diagnostic test discrimination (good, reasonable, poor), and reader error levels (low, medium, high). For each combination, 1,000 simulations were performed, and variability measures (percentage agreement, Cohen's kappa, Prevalence-Adjusted Bias-Adjusted Kappa (PABAK), Krippendorff's alpha, and Gwet's AC coefficient) were calculated, along with sensitivity and specificity.

Results

The study showed that increased sample size consistently produced more precise variability estimates across all scenarios. Percentage agreement consistently showed the highest values among the variability measures. PABAK and Gwet's AC coefficient demonstrated greater stability and less sensitivity to condition prevalence compared to Cohen's kappa and Krippendorff's alpha, which showed more variable performance. As diagnostic test discrimination decreased and reader error increased, all variability measures showed a decline.

Conclusion

These findings show the importance of considering different factors in assessing interobserver variability in diagnostic imaging tests. Different variability measures are affected in distinct ways by participant sample size, condition prevalence, diagnostic test discrimination, and reader error levels. By providing guidance on designing interobserver variability studies, future studies can be improved, providing more accurate information on the reliability of diagnostic imaging tests, leading to better patient care.

43: Simulation-Based Optimization of Adaptive Designs using a Generalized Version of Assurance

Pantelis Vlachos¹, Valeria Mazzanti¹, Boaz Adler²

¹Cytel Inc, Switzerland

²Cytel Inc, USA

The power of cloud computing is utilized to create a tool that collects information from different parts of the clinical development team (clinical, operations, commercial etc) and with the statistician at the driver seat seeks and proposes designs that optimize a clinical study with respect to sample size, cost, duration and power. The optimization is performed using a generalized assurance measure that takes into account all trial possible scenarios with respect to treatment effect, control response, enrollment, dropouts etc. Furthermore, this tool can be used to communicate and update information to the trial team in real time, considering (possibly) changing target objectives. Case studies of actual adaptive trials will be given.

44: Evaluating the Impact of Outcome Delay on Adaptive Designs

Aritra Mukherjee¹, Michael J. Grayling², James M. S. Wason¹

¹Population Health Sciences Institute, Newcastle University

²Johnson and Johnson

Background Adaptive designs (AD) are a broad class of trial designs that allow pre-planned modifications to be made to a trial as patient data is accrued, without undermining its validity or integrity. ADs can lead to improved efficiency, patient-benefit, and power of a trial. However, these advantages may be affected adversely by a delay in observing the primary outcome variable. In the presence of such delay, a choice must be made between (a) pausing recruitment until requisite data is accrued for the interim analysis, leading to longer trial completion period; or (b) continuing to recruit patients, which may result in a large number of participants who do not benefit from the interim analysis. In the latter case, little work has investigated the size of outcome delay that results in the realised efficiency gains of ADs being negligible compared to classical fixed-sample alternatives. Our study covers different kinds of ADs and the impact of outcome delay on them.

Methods We assess the impact of delay on the expected efficiency gains of an AD by estimating the number of pipeline patients being recruited in the trial under the assumption that recruitment is not paused while we await treatment outcomes. We assume different recruitment models to suitably adjust for single- or multi-centred trials. We discuss findings for two-arm group-sequential designs as well as multi-arm multi-stage designs. Further, we focus on sample size re-estimation (SSR), a design where the variable typically optimized to characterise trial efficiency is not the expected sample size (ESS).

Results and conclusions Our results indicate that if outcome delay is not considered at the planning stage of a trial, this can translate to much of the expected efficiency gains being

Abstracts of Contributed Posters

lost due to delay. The worst affected designs are typically those with early stopping, where the efficiency gains are assessed through a reduced ESS. SSR can also suffer adversely if the initial sample size specification was largely over-estimated.

Finally, in light of these findings, we discuss the implications of using the ratio of the total recruitment length to the outcome delay as a measure of the utility of different ADs.

Wednesday Posters at ETH

Wednesday, 2025-08-27 09:00 - 10:30, ETH, UG hall

1: Exploring the Exposome Correlated with Body Mass Index in Adolescents: Findings from the 2014-2015 and 2022-2023 KNHANES

Hye Ah Lee¹, Hyesook Park²

¹Clinical Trial Center, Ewha Womans University Mokdong Hospital, Seoul, Republic of Korea

²Department of Preventive Medicine, Graduate Program in System Health Science and Engineering, Ewha Womans University, Seoul, Republic of Korea

Background: To identify multifaceted features correlated with body mass index (BMI) in adolescents, we conducted an exposome-wide association study (ExWAS) using data from the Korea National Health and Nutrition Examination Survey (KNHANES), a nationally representative survey.

Methods: To obtain robust findings, we constructed a multi-year dataset covering two study periods (2014-2015 and 2022-2023). Adolescents aged 12 to 18 years with complete BMI data were included, while those dieting for weight loss or health conditions were excluded. This resulted in 941 participants from the 2014–2015 dataset and 637 from the 2022–2023 dataset. Approximately 130 features derived from questionnaires, health examinations, and dietary surveys were analyzed. Standardized BMI (zBMI) was used as the outcome, and ordinal or numeric features were standardized by sex and age using mean and standard deviation. ExWAS was performed through survey-design-based linear regression, adjusting for sociodemographic features. Additionally, pairwise relationships between features were assessed using a mixed graphical model (MGM) network.

Results: In the 2022–2023 dataset, 20.2% of boys and 15.0% of girls were classified as obese. Of the approximately 130 exposomes, 13 features in boys and 9 features in girls were selected as features correlated with BMI. Boys who perceived themselves as unhealthy or considered their body shape as fat also had higher BMI. zBMI was positively correlated with alanine aminotransferase (ALT), white blood cell (WBC), platelets, systolic blood pressure (SBP), total cholesterol, and triglyceride (TG) and negatively correlated with high density lipoprotein cholesterol (HDL-C). These trends were also observed in the 2014–2015 dataset. Among girls, zBMI was positively correlated with ALT, WBC, SBP, and TG and negatively correlated with HDL-C. Girls who perceived their body shape as fat had higher BMI, consistent with findings from the 2014–2015 dataset. Notably, in the 2022–2023 dataset, girls who reported

suicidal thoughts had higher BMI. In the MGM network analysis, ALT, WBC, and HDL-C were directly correlated with zBMI across all datasets, regardless of sex.

Conclusion: In adolescents, metabolic indices showed a clear correlation with BMI, and in addition to the commonly considered metabolic indices, ALT and WBC were directly correlated. Furthermore, subjective body shape perception, as assessed through questionnaires, was significantly correlated with BMI.

2: Flexible Statistical Modeling of Undernutrition among under-Five Children in India

Shambhavi Mishra

UNIVERSITY OF LUCKNOW, India

Background Childhood undernutrition has an irreversible impact on the physical as well as mental development of the child. Nutrition-related factors were responsible for about 35% of child deaths and 11% of the total global disease burden. This health condition continues to be a major public health issue across the globe.

Methods Three standard indices based on anthropometric measurements viz. weight and height, that describe nutritional status of children are: height-for-age (stunting), weight-for-age (underweight) and weight-for-height (wasting). Z-scores have been computed on the basis of appropriate anthropometric indicators (weight & height) relative to the WHO International reference population for the particular age. This paper utilises unit-level data on under-five children of India from the NFHS-5, 2019-2021 to find out factors which exert a differential impact on the conditional distribution of the outcome variable. A class of models that allow flexible functional dependence of an outcome variable on covariates by using non-parametric regression have been applied to determine possible factors causing undernutrition. This study also fits a Bayesian additive quantile regression model for the provision of a complete picture of the relationship between the outcome variable and the predictor variables on different desired quantiles of the response distribution. Different types of quantile regression models were fitted and compared according to each Deviance Information Criteria (DIC) for determination of the best model among them.

Results Maternal characteristics like nutrition, education showed significant impact on child's nutritional status, consistent with the findings of other studies. Child' s age and Mother's nutrition were among the continuous factors exerting non-linear effect on stunting, with mother's BMI showing maximum effect size at lower end of the distribution. Also it could be

seen that maximum number of covariates were found significant for severe undernutrition, indicating that differential effect of predictors on the conditional distribution of the outcome variables.

Conclusions Although widely applicable, logistic regression model enables the researcher to have an idea of the determinants of undernutrition, providing only a preliminary basis. To study variables such as nutritional status of children, where lower quantiles are of main interest, focus should be on how factors affect the entire conditional distributional of the outcome variable taken as is rather than summarizing the distribution at its mean. This can be achieved by applying quantile regression modeling. An extension to it further enables to non-parametrically estimate the linear or potentially non-linear effects of continuous covariates differentially on the outcome using penalized splines.

3: Comparison of Deep Learning Models with Different Architectures and Training Populations for ECG Age Estimation: Accuracy, Agreement, and CVD Prediction

Arya Panthalanickal Vijayakumar¹, Tom Wilsgaard¹, Henrik Schirmer^{2,3}, Ernest Diez Benavente⁴, René van Es⁵, Rutger R. van de Leur⁵, Haakon Lindekleiv⁶, Zachi I. Attia⁷, Francisco Lopez-Jimenez⁷, David A. Leon⁸, Olena Iakunchykova⁹

¹Department of Community Medicine, UiT The Arctic University of Norway, Norway

²Akershus University Hospital, Lørenskog, Norway

³Institute of Clinical Medicine, Campus Ahus, University of Oslo, Norway

⁴Department of Experimental Cardiology University Medical Center Utrecht, The Netherlands

⁵Department of Cardiology University Medical Center Utrecht, The Netherlands

⁶Department of Radiology, University Hospital of North Norway

⁷Mayo Clinic College of Medicine, Rochester, MN, USA

⁸Department of Noncommunicable Disease Epidemiology, London School of Hygiene & Tropical Medicine, London, United Kingdom

⁹Department of Psychology, University of Oslo, Norway

Background Several convolutional neural networks (CNNs) have been developed to estimate biological age based on 12-lead electrocardiograms (ECG) - ECG age. This new biomarker of cardiac health can be used as a predictor of cardiovascular disease (CVD) and mortality. Before implementation into clinical practice, it is crucial to compare proposed CNN models used to estimate ECG age to assess their accuracy, agreement, and predictive abilities on external sample.

Methods We used 7,108 participants from the Tromsø Study (2015-16) to compare ECG ages estimated with three different previously proposed CNNs. CNNs differed by model architecture and/or population that they were trained and tested on. We calculated mean absolute error (MAE) for each CNN. Agreement was assessed using Pearson and intraclass correlation coefficients (ICC), and Bland-Altman (BA) plots. The predictive abilities of each ECG age or δ -age (difference between ECG age and chronological age) were assessed by the concordance index (C-index) and hazard ratios (HRs) from Cox proportional hazards models for myocardial infarction (MI), stroke, CVD mortality, and all-cause mortality, with and without adjustment for traditional risk factors.

Results All three CNNs had fairly close MAEs (6.82, 7.82, and 6.42 years) and similar Pearson correlation coefficients with chronological age (0.72, 0.71, and 0.73, respectively). Visual agreement using BA plots was good, and the ICC indicated good agreement (0.86; 95% CI: 0.86, 0.87). The multivariable adjusted HRs for MI and total mortality were strongest for δ -age₁ (HR 1.36 (1.11, 1.67) and 1.27 (1.08, 1.50), respectively, while HRs for stroke and CVD mortality were strongest for δ -age₂ (HR 1.45 (1.17, 1.80) and 1.48 (1.07, 2.05), respectively. The 6-year survival probability predictions showed excellent agreement among all δ -ages for all outcomes in terms of both BA plots and ICC. The C-index values showed no significant difference between pairwise combinations of models with ECG age₁, ECG age₂, or ECG age₃ for all outcomes.

Conclusion We observed good agreement between ECG ages estimated by three different CNNs in terms of accuracy, agreement, and predictive ability. We did not identify that one CNN for ECG age is superior over another for prediction of CVD outcomes or death in the Tromsø Study.

4: Systematic Review and Real Life-Oriented Evaluation on Methods for Feature Selection in Longitudinal Biomedical Data

Alexander Gieswinkel^{1,2,3}, **Gregor Buch**^{1,3}, **Gökhan Gül**^{1,4}, **Vincent ten Cate**^{1,3,4},
Lisa Hartung², **Philipp S. Wild**^{1,3,4,5}

¹Preventive Cardiology and Preventive Medicine, Department of Cardiology, University Medical Center of the Johannes Gutenberg University Mainz, 55131 Mainz, Germany

²Institute of Mathematics, Johannes Gutenberg University Mainz, 55128 Mainz, Germany

³German Center for Cardiovascular Research (DZHK), partner site Rhine Main, 55131 Mainz, Germany

⁴Clinical Epidemiology and Systems Medicine, Center for Thrombosis and Hemostasis, University Medical Center of the Johannes Gutenberg University Mainz, 55131 Mainz, Germany

⁵Institute of Molecular Biology (IMB), 55131 Mainz, Germany

Background High-dimensional omics data are increasingly commonly available for longitudinal cohort studies since biochemical technology is improving. Supervised feature selection based on biomedical data from multiple time points is often required. However, an overview of existing methods for this setting is lacking, which motivated a systematic review and evaluation of this area.

Methods A systematic search of statistical software was conducted to identify relevant methods. The Comprehensive R Archive Network (CRAN) was examined via the R package '*packagefinder*' with a search query containing relevant keywords. Eligible software was characterised by manually screening the package descriptions, and through computational testing with a fix application example. An ADEMP-designed simulation study was conducted to evaluate the identified methods in real-world scenarios, considering varying sample sizes, predictors, time points and signal-to-noise ratios. Only frequentist implementations with given default settings were included for a fair comparison. The estimated true positive rate (eTPR) and estimated false discovery rate (eFDR) were chosen as performance measures.

Results Of 21,528 accessible packages on CRAN, 324 packages with matching keywords in the descriptions were extracted by the search query. Screening of the descriptions identified 45 packages that were then tested in R, leading to 14 packages. Six packages were based on mixed effects models ('*buildmer*', '*rpql*', '*splmm*', '*alqrfe*', '*plsommLasso*'; '*glmmLasso*'), five on generalized estimating equations ('*sgee*', '*LassoGEE*', '*geeVerse*'; '*PGEE*', '*pgee.mixed*'), two methods were built on Bayesian frameworks ('*sparsereg*'; '*spikeSlabGAM*') and one package was modelling time series ('*midasmf*'). All implementations were able to process continuous outcomes, while only four supported binary outcomes. A total of N=8 frequentist methods with sufficient default settings were considered in the simulation study.

The packages '*buildmer*' and '*plsommLasso*' consistently demonstrated an eTPR exceeding 80% while maintaining the eFDR under 20%, across various signal-to-noise settings. By comparison, all other methods underperformed in jointly evaluating both performance metrics. '*splmm*' achieved similar eFDR but yielded lower eTPR, whereas '*geeVerse*' showed an opposite trend. In contrast, both '*rpql*' and '*alqrfe*' failed to select any variables.

Conclusions Supervised feature selection in longitudinal biomedical data can be performed using a variety of methods. The majority of the available statistical software is based on frequentist techniques, while Bayesian procedures represent a minority. Alternative concepts like tree-based methods are notably absent. No evidence of superiority was found for modern selection techniques such as penalized regression ('*plsommLasso*') over traditional approaches like stepwise regression ('*buildmer*') for feature selection in longitudinal data.

5: How Does Acute Exposure to Environmental Factors Relate to Stroke Characteristics Such as Stroke Type, Severity, and Impairments?

Bohan Zhang¹, Andy Vail¹, Craig Smith², Amit Kishore², Matthew Gittins¹

¹University Of Manchester, United Kingdom

²Manchester Centre for Clinical Neuroscience, Manchester, United Kingdom

Background and Aims: The overall aim of this subject is to better understand the association between acute exposure to environmental factors such as ambient air pollution and temperature and stroke characteristics. We aim to focus on the short-term acute effects associated with same-day or up to 30 days before stroke. Specifically, I will look into Stroke Counts and Stroke Severity using non-identifiable patients data from SSNAP from Manchester Stroke Units.

Methods We may employ a cohort (or case-control design), where the cohort is all stroke patients within Greater Manchester/Salford, their exposure is the exposure leading up to stroke, and their outcomes are the post-stroke characteristics. It's more likely to be a cohort study, but the case-control might help deal with some of the selection issues, i.e. the group is defined by being a stroke patient. Rather than being identified before and following up to see if they become a stroke patient or not.

We will employ methods to model the lagged effects of air pollution such as the lag stratified model (where days are grouped and average exposure is modelled), and the distributed lag models (where polynomial functions are applied to represent the 30 days).

Results The results are still under way and will be presented on ISCB.

Conclusion From some literature on other diseases it often seen that extreme condition in Environmental Exposure is highly likely to lead to a worse outcome. But the result is still on the way

6: Improving TBI Prognosis in Developing World: A Machine Learning-Based AutoScore Approach to Predict Six-Month Functional Outcome

Vineet Kumar Kamal¹, Deepak Agrawal²

¹AIIMS, Kalyani

²AIIMS, New Delhi

Background Traumatic brain injury (TBI) presents a significant challenge in predicting long-term functional outcomes due to its complex nature and variability among patients. Accurate prognostic tools are essential for clinicians to guide treatment decisions and set realistic expectations for recovery. To address this, AutoScore employs a machine learning-based approach that automates the generation of clinical scores, facilitating the prediction of outcomes. This study aims to develop, validate and to see clinical utility of a prognostic model to accurately predict six-month functional outcomes in severe/moderate, adult TBI patients, enhancing risk stratification.

Methods This retrospective cohort study included 1,085 adult patients with TBI from a public, tertiary care, level-1 trauma center in India. We considered a total of 72 demographic, clinical, secondary insults, CT and lab variables from admission to first discharge. We developed the AutoScore framework, consisting of six distinct modules: variable ranking, variable transformation, score derivation, model selection, score fine-tuning, and model evaluation. We divided the whole dataset randomly into (0.7, 0.1, 0.2) to develop, parameter tuning/validation, and testing. The predictive performance of the AutoScore framework was evaluated using various metrics, including receiver operating characteristic (ROC) curves, calibration curves, brier score, and decision curves for clinical utility analysis. All the analyses were performed using R software v.4.3.3.

Results The AutoScore model identified only four key risk predictors: motor response at discharge, verbal response at discharge, motor response at the time of admission, and eye-opening response at discharge, with higher scores indicating an increased risk of an unfavorable six-month outcome in TBI patients. The final model achieved an AUC of 0.93 (95% CI: 0.88–0.98) on the validation set and 0.81 (95% CI: 0.76–0.86) on the test set, demonstrating strong predictive performance. Brier score was 0.14 and graphical plot, and observed-to-expected ratio 0.978 suggested that the model was well-calibrated in test data. **Our model was useful in the 0.0–0.6 threshold range** (offers better net benefit). The predicted risk increased steadily with the total score, as depicted in the probability plot, with patients scoring above 75 exhibiting near-certain risk of an unfavorable outcome.

Conclusion The AutoScore-based prognostic model demonstrated strong predictive performance for six-month functional outcomes in moderate-to-severe TBI patients using only four key predictors. These findings suggest that the model could serve as a valuable tool for clinicians in early risk assessment and decision-making. Further validation in diverse populations with recent data is warranted to confirm its generalizability and clinical applicability.

7: Predictive Risk Index for Poor Cognitive Development Among Children Using Machine Learning Approaches

Anita Kerubo Ogero¹, Patricia Kipkemoi¹, Amina Abubakar^{1,2,3}

¹Aga Khan University, Nairobi, Kenya, Institute for Human Development, Aga Khan University, P.O. BOX 30270-00100, Nairobi, Kenya

²Centre for Geographic Medicine Research Coast, Kenya Medical Research (KEMRI), P.O Box 230-80108, Kilifi, Kenya

³Department of Psychiatry, University of Oxford, Warneford Hospital, Warneford Ln, Oxford OX37JX, United Kingdom

Poor cognitive development in early childhood is a major global concern, with over 200 million children failing to reach their developmental milestones due to factors like malnutrition and poverty - particularly in low- and middle-income countries. Cognitive abilities established during childhood are critical determinants of a child's future academic and socio-economic outcomes. Despite extensive research on socio-demographic, environmental and nutritional influences on cognitive development, there remains a gap in developing a predictive risk index tailored for resource-constrained settings. Early identification of at-risk children is essential to enable timely interventions and inform policy. In this study, we propose to develop and validate a risk index for poor cognitive development among children using advanced machine-learning techniques. Secondary data from approximately 7,000 children, assessed with the Raven's Progressive Matrices (RPM), will be analysed; cognitive development is classified into no-risk, low-risk, and high-risk groups based on age-adjusted percentile scores. Predictor variables integrate socio-demographic factors (e.g., parental education, socioeconomic status) and nutritional indicators (e.g., anthropometric measurements such as height, weight, head circumference, and derived indices like weight-for-age and height-for-age z-scores). Our analytic framework integrates several methods including logistic regression, Random Forest, Support Vector Machines, Artificial Neural Networks, and Extreme Gradient Boosting. Data preprocessing involves feature selection via Recursive Feature Elimination (RFE) and dimensionality reduction using Principal Component Analysis (PCA). Decision thresholds will be optimised through the Receiver Operating Characteristic (ROC) curve analysis and Youden's Index to balance sensitivity and specificity. Key risk factors significantly associated with poor cognitive development will be identified, forming the basis for a validated risk index. The risk index will be assessed for predictive accuracy and generalisability. The developed risk index will represent a significant advancement in the early identification of children at risk for poor cognitive development in low-resource environments. Findings may inform policy decisions and the development of digital tools, such as mobile applications, for real-time cognitive risk assessment. Moreover, this tool holds promise for improving long-term developmental outcomes by optimising resource allocation and enabling targeted interventions.

8: Development and Validation of a Model to Predict Ceiling of Care in COVID-19 Hospitalised Patients

Natàlia Pallarès Fontanet¹, Hristo Inouzhe², Jordi Cortés³, Sam Straw⁴, Klaus K Witte⁴, Jordi Carratalà⁵, Sebastià Videla⁶, Cristian Tebé¹

¹Biostatistics Support and Research Unit, Germans Trias i Pujol Research Institute and Hospital (IGTP), Badalona, Spain

²Basque Center for Applied Mathematics, BCAM, Bilbao, Spain

³Department of Statistics and Operations Research, Universitat Politècnica de Catalunya/BarcelonaT-ech, Barcelona, Spain

⁴Leeds Institute of Cardiovascular and Metabolic Medicine, University of Leeds, Leeds, UK

⁵Department of Infectious Diseases, Bellvitge University Hospital, Barcelona, Spain

⁶Clinical Research Support Area, Department of Clinical Pharmacology, Germans Trias i Pujol University Hospital, Badalona, Spain

Background Therapeutic ceiling of care is the maximum therapeutic effort to be offered to a patient based on age, comorbidities, and the expected clinical benefit in relation to resource availability. COVID-19 patients with and without an assigned ceiling of care at hospital admission have different baseline variables and outcomes. Analysis of hospitalised COVID-19 subjects should be stratified by ceiling of care to avoid bias, but there are currently no models to predict their ceiling of care. We aimed to develop and validate a clinical prediction model to predict ceiling of care at hospital admission.

Methods The data used to develop the model came from an observational study conducted during four waves of COVID-19 in 5 centres in Catalonia. Data were sampled 1000 times by bootstrapping. For each sample, a logistic regression model with ceiling as outcome was fitted using backward elimination. Variables retained in more than 95% of the models were candidates for the final model. Alternative variable selection methods such as Lasso, CART, and Boruta were also explored. Discrimination was assessed by estimating the area under the ROC curve and the Brier Score, and calibration by comparing observed versus expected probabilities of ceiling of care by deciles of predicted risk. The final model was validated internally, and externally using a cohort from the Leeds Teaching Hospitals NHS Trust.

Results A total of 5813 patients were included in the development cohort, of whom 31.5% were assigned a ceiling of care on admission. A model including age, COVID-19 wave, chronic kidney disease, dementia, dyslipidaemia, heart failure, metastasis, peripheral vascular disease, chronic obstructive pulmonary disease, and stroke had excellent discrimination (AUC 0.898 [0.889; 0.907]; Brier Score 0.113) and calibration (slope of the regression line between observed and predicted $\beta=1.01$ [0.94; 1.08]) in the whole cohort and in subgroups of interest. External validation on the Leeds Teaching Hospitals cohort also showed good performance

(AUC 0.934 [0.908; 0.959]; Brier Score 0.110; $\beta=0.98$ [0.80; 1.17]).

Conclusions Ceiling of care can be predicted with great accuracy from baseline information available at hospital admission. Cohorts without information on ceiling of care could use our model to estimate the probability of ceiling of care. This model, combined with clinical expertise, may be valuable in future pandemics or emergencies requiring time-sensitive decisions about life-prolonging treatments, but further evaluation outside of COVID-19 is needed.

9: Transformer Models for Clinical Prediction – Investigation of BEHRT in UK Biobank and Prediction Assessment under Different Scenarios

Yusuf Yildiz, Goran Nenadic, Meghna Jani, David A. Jenkins

The University of Manchester, United Kingdom

Background Transformer-based Large Language models (LLMs) like BEHRT¹ have shown potential in modelling Electronic Health Records to predict future instances. These models can represent patient histories by including structured (diagnoses) and unstructured data (doctor notes)². BEHRT showed superior performance over the state-of-the-art models at the time its developed using a large primary care data set. However, it's unclear if such model and high accuracy can be achieved for other real-world datasets i.e. hospital data. Developing LLMs requires selecting various decisions like data split strategies, medical terminology selection and parameters. Parameter choices have been shown to impact model performance, stability and generalisability, but it's unclear the extent this also hold for LLMs. This study aims to implement the BEHRT architecture in the UK Biobank and identify challenges of implementing this model into different dataset. The secondary aim is to assess the impact of parameter choices on prediction performance.

Methods This study uses UK Biobank data. To capture key features of patient histories, embeddings are created using diagnoses and age at diagnosis. BEHRT workflow included pre-training with masked language modelling (MLM) and fine-tuning for next-diseases prediction across different time frames. Prediction performance was evaluated using Average Precision Score and AUROC. Initially, the original study is replicated using UK Biobank to assess the impact of dataset variability. Subsequently, the model's performance was evaluated to assess the effects of different medical terminologies (ICD10 and CALIBER phenotyping) and data splits.

Results/Conclusion:

Results showed that decisions that we make while we develop these models using different datasets effects the performance of the model. Our replicated BEHRT model did not achieve as hight predictive performance as performance metrics as the original. Terminologies with bigger vocabularies showed worse performance. Complete separation of the MLM and fine-tuning data resulted with worse performed model. However, most developed models use complete dataset for pre-training and therefore are likely to exhibit overly optimistic performance.

Also, more rigorous, definitive framework and assessment workflow is needed for LLM development in clinical prediction. Especially clinical usefulness of these model should be examined. Reporting guidelines, like TRIPOD-LLM³ should be used for transparent model development.

Further work is needed on time-to-event analysis, censoring adjustment, transparent decision-making and computational costs adjustment for better integration into clinical prediction.

10: What is the Best Way to Analyse Ventilator-Free Days?

Laurent Renard Triché^{1,2,3}, Matthieu Jabaudon^{1,2}, Bruno Pereira⁴, Sylvie Chevret^{2,5}

¹Department of Perioperative Medicine, CHU Clermont-Ferrand, Clermont-Ferrand, France

²iGReD, INSERM, CNRS, Université Clermont Auvergne, Clermont-Ferrand, France

³ECSTRRA Team, IRSL, INSERM UMR1342, Université Paris Cité, Paris, France

⁴Biostatistics Unit, Department of Clinical Research, and Innovation (DRCI), CHU Clermont-Ferrand, Clermont-Ferrand, France

⁵Department of Biostatistics, Hôpital Saint-Louis, AP-HP, Paris, France

Introduction Ventilator-free days (VFDs) are a composite outcome increasingly used in critical care research, reflecting both survival and mechanical ventilation duration. However, inconsistencies exist in the models used to analyse VFDs. Some researchers evaluate VFDs as a count, primarily using the Mann-Whitney statistics, while others consider them as a time-to-event outcome, where survival is a competing risk for extubation. Alternative approaches such as the multi-state model and the win ratio warrant investigation.

This study aimed to evaluate different statistical models to determine the best approach for analysing VFDs.

Methods First, a clinical trial dataset (LIVE study, NCT02149589) was used to apply different statistical models to analyse VFDs. Then, 16 datasets of 300 individuals were simulated

with 3,000 independent replications, comparing a control group with an intervention strategy by varying survival rates and ventilation durations derived from exponential distributions. The simulated data were analysed using the same statistical methods, and statistical power and type I error rates were compared between different models.

Eleven statistical methods were evaluated, including the Mann-Whitney test, the zero-inflated negative binomial model, the negative binomial hurdle model, the zero-inflated Poisson model, the Poisson hurdle model, the log-rank test, the Gray test, the cause-specific hazard model, the Fine-Gray model, the multistate Markov model, and the win ratio.

In addition, three sensitivity analyses were performed by adjusting the survival rates and/or ventilation durations in the control group.

Results In the LIVE study, almost all methods identified a significant association between VFDs (or related measures) and the patient groups, except for the count submodels, the log-rank test, and the cause-specific hazard model for the survival.

For the simulated data, the 28-day mortality rate was set at 20% and the mean duration of ventilation at 15 days for the control group. Most statistical methods effectively controlled the type I error rate, although exceptions included the zero-inflated and hurdle Poisson/negative binomial count sub-models and the cause-specific Cox regression model for survival. Statistical methods had variable power to detect survival benefits and effects on duration of ventilation, with the time-to-event approach and the win ratio generally having the highest power.

The sensitivity analyses found similar results.

Conclusion The time-to-event approach and the win ratio were more appropriate than the count-based methods to analyse the VFDs and may be extended to other free-days outcomes. Simulation should be recommended for power calculation and sample size estimation rather than a simplified formula.

11: Compare Estimation and Classification Performances of Statistical Shrinkage Methods Ridge Regression, Lasso Regression, and Elastic Net Regression

Gamze Ozen, Fezan Mutlu

Eskisehir Osmangazi University, Medical Faculty, Department of Biostatistics, Eskisehir, Turkey

Introduction Advances in data science indicate the need to improve the reliability of regression model estimation when the number of independent variables exceeds the number of observations in multidimensional datasets. Such a dataset's multicollinearity causes the accuracy of prediction models to be reduced. This study aims to assess the performance of Ridge, Lasso, and Elastic net regression methods in the case of multicollinearity and multidimensional datasets.

Method Performance of three regression methods where Ridge, Lasso, and Elastic Net is verified by data stimulation that Elastic Net method exhibits superiority to all the strongly correlated variables into the model over Ridge and Lasso methods. Models are applied to the dataset containing the serum miRNA in large cohorts to identify the miRNAs that can be used to detect breast cancer in the early stage (Shimomura et al., 2016).

Results Data simulations verify that Elastic Net regression produces better results with an accuracy of 0.963 when the data is high-dimensional and has strong multicollinearity. A determination of breast cancer by miRNAs shows that Elastic Net can use classification with 96% accuracy.

Conclusion The findings suggest that statistical Shrinkage Methods such as Ridge Regression, Lasso Regression, and Elastic Net Regression are reliable and useful for prediction and classification research on linear and logistic models. This study suggests that Statistical Shrinkage Methods may be enhanced in health science to generate stronger models.

12: Defining Harm in Settings with Outcomes that are not Binary

Amit Sawant, Mats Stensrud

EPFL, Switzerland

The increasing application of automated algorithms in personalised medicine necessitates that algorithm recommendations do not harm patients, in accordance with the Hippocratic maxim of “Do no harm.” A formal mathematical definition of harm is essential to guide these algorithms in adhering to this principle. A counterfactual definition of harm has been previously proposed, which asserts that a treatment is considered harmful if there exists a non-zero probability that the potential outcome under treatment for an individual is worse than the potential outcome without treatment. Existing literature on counterfactual harm has primarily focused on binary treatments and outcomes. This study aims to illustrate that in scenarios involving multiple treatments and multi-level outcomes, the counterfactual definition of harm can result in intransitivity in the ranking of treatments. Specifically, we

analyse three treatments—A, B, and C—for a particular disease. We demonstrate that treatment B is less harmful than treatment A, treatment C is less harmful than treatment B, yet treatment C is more harmful than treatment A in direct comparison, if we follow the counterfactual definition. Our example highlights that the intuitive concept of counterfactual harm in binary settings does not extend to scenarios involving more than two treatments and outcomes. On the other hand, an interventionist definition of harm in terms of utility circumvents the issue of intransitivity.

13: Brier Pseudo-Observation Score for Selecting a Multiplicative, an Additive or an Additive-Multiplicative Hazards Regression Model

François Lefebvre¹, Roch Giorgi²

¹Groupe méthode en recherche clinique, service de santé publique, Hôpitaux universitaires de Strasbourg, Strasbourg, France

²Aix Marseille Univ, APHM, Inserm, IRD, SESSTIM, ISSPAM, Hop Timone, BioSTIC, Marseille, France

Background In survival analysis, data can be modelled in different ways: with the Cox model, with an additive hazards model, as the Aalen's model or with an additive-multiplicative model, as the Cox-Aalen model. Covariates act on the baseline hazard multiplicatively in the first model, additively in the second and some of these act multiplicatively, others additively in the third. Correct modelling of the covariates requires knowledge of its effect on the baseline hazard, which is rarely known *a priori*. The pseudo-observations has been used in the evaluation of the impact of a covariate on survival outcomes, in addition to the verification of the assumptions inherent in the Cox (proportional hazards, log-linearity) and the Aalen (linearity) models [1]. Nowadays, they do not permit to know which one of the multiplicative, additive or additive-multiplicative model is the more appropriate for a particular survival dataset. The aim of this study is to propose a method for selecting a multiplicative, an additive or an additive-multiplicative hazards regression model adapted to the survival data-generating mechanism.

Methods We propose to use the Brier pseudo-observation score defined by Perperoglou [2] as the mean of the square difference of the pseudo-observations and the survival estimates obtained using a regression model. Therefore, for each type of regression model, Brier pseudo-observation score can be computed and compared to each other. Since the Brier pseudo-observation score is analogous to the mean square error of prediction, the lower the score the better the model. In order to reduce the risk of overfitting, the model parameters were estimated for each individual using Jackknife. Performance of this approach was

assessed in simulation studies comparing Brier pseudo-observation score obtained with a multiplicative, an additive and an additive-multiplicative model, in situations in which survival data-generating mechanism was either multiplicative, additive or had both effect.

Results This measure selected the model used to generate the data in over 80% in most of the scenarios considered. The utilisation of this approach was exemplified by an epidemiological example of female breast cancer with the objective of ascertaining the impact of nodal status, age and tumour size on the baseline hazard.

Conclusion This method has been demonstrated to achieve optimal performance in the selecting the hazards regression model adapted to the data-generating mechanism.

[1] M. Pohar Perme, K. Andersen. Statistics in Medicine, 27, 2008, 5309–5328.

[2] A. Perperoglou, A. Keramopoullos, H. C. van Houwelingen. Statistics in Medicine, 26, 2007, 2666–2685.

14: Bayesian Spatio-Temporal Analysis of the COVID-19 Pandemic in Catalonia

Pau Satorra, Cristian Tebé

Biostatistics Support and Research Unit, Germans Trias i Pujol Research Institute and Hospital (IGTP), Spain

Introduction The COVID-19 pandemic posed an unprecedented challenge to public health systems worldwide. The spread of the pandemic varied in different geographical regions, even at the level of small areas. This study investigates the spatio-temporal evolution of COVID-19 cases and hospitalisations in the different basic health areas (ABS) of Catalonia during the pandemic period (2020-2022). Additionally, it assesses the impact of demographic and socio-economic factors, as well as vaccination coverage, on infection and hospitalisation rates at an ABS level.

Methods Data were obtained from the official open data catalogue of the Government of Catalonia. Bayesian hierarchical spatio-temporal models were used, estimated with Integrated Nested Laplace Approximation (INLA). Demographic and socio-economic ABS variables were included in the models to assess its role as risk factors for cases and hospitalisations. Full ABS vaccination coverage was also incorporated to assess its effect. All analyses were performed using the R statistical program.

Results During the study period, a cumulative total of 2,685,568 COVID-19 cases and 144,550 hospitalisations were reported in Catalonia, representing a 35% and a 1.89% of the total population, respectively. The estimated spatial, temporal and spatio-temporal relative risks (RR) were visualized through maps and plots, identifying high-risk (hotspots) and low-risk (coldspots) areas and weeks. These results were presented in an interactive R-shiny application: https://brui.shinyapps.io/covidcat_evo/. Urban areas had a higher risk of cases (RR: 5%, CI95%: 2-9%) and hospitalisations (RR: 17%, CI95%: 10-25%). Higher socio-economic deprivation index was associated with an increased hospitalisation risk (RR: 19%, CI95%: 17-22%). Finally, a higher full vaccination coverage in the ABS was associated with a reduced risk of cases (RR: 12%, CI95%: 5-18%) and hospitalisations (RR: 17%, CI95%: 2-32%) during the fourth and fifth pandemic waves.

Conclusion This study provides a comprehensive study to understand the COVID-19 pandemic across the territory of Catalonia at the small area level, revealing the spatial, temporal and spatio-temporal patterns of the disease. Urban areas had a higher risk of COVID-19 cases and hospitalisations, socio-economic deprivation increased hospitalisations, and full vaccination was protective against cases and hospitalisations during specific pandemic waves. These findings offer valuable insights for public health policymakers to design targeted interventions against future infectious disease threats.

15: A Simulation Study of Bayesian Approaches to Spatial Modelling Using the Besag-York-Mollie Model

Hollie Hughes, David Hughes

Department of Health Data Science, University of Liverpool, United Kingdom

Background/Introduction Spatial modelling can be a useful tool for analysing patterns and relationships in data to indicate how events might be spatially related. It is known that neighbouring areas tend to be more strongly correlated and share similar characteristics than distant areas when modelling and mapping data, creating a spatial autocorrelation problem. Spatial models have been successfully developed to account for this autocorrelation problem in areal data, allowing patterns to be successfully modelled. However, to do this in a Bayesian framework, the Markov Chain Monte Carlo (MCMC) method can often be computationally expensive, particularly in larger spatial datasets. Therefore, many researchers opt for the Integrated Nested Laplace Algorithm (INLA) approach for computational savings. We suggest an alternative using approximate Mean Field Variational Bayes (MFVB) algorithms to decrease the computational burden as the INLA approach does, whilst potentially sustaining accuracy that is promised through the MCMC approach.

Method We provide a comparison of the MCMC, INLA and MFVB approach to the Besag-York-Mollie (BYM) model which is commonly used for spatial modelling to account for spatial dependencies. We conducted a simulation study to compare the performance of the three approaches to fitting the BYM model to spatially structured data on incidence of Depression. Synthetic datasets were generated under the BYM model specification outlined in Morris (2019), incorporating both spatially structured and unstructured random effects (1).

Each method was implemented using standard Bayesian modelling tools in R: INLA via the R-INLA package, MCMC using Stan, and MFVB using Stan's Variational Bayes options. We assessed computational efficiency, and accuracy for each method by comparing posterior estimates against the true simulated values and measuring time taken to fit each model. Accuracy of results were assessed both in terms of distributional similarity and accuracy of point estimates.

Results Results will include comparisons of accuracy and performance metrics including measures comparing the ground truth with MCMC results and computation time for each model. The results will be summarised across multiple simulated datasets to evaluate consistency and robustness. Evaluation is ongoing but full results will be presented at the conference.

Conclusion This simulation study may indicate the usefulness of the MFVB approach as an alternative to the MCMC approach with the potential of being substantially as accurate when simulated values are known, alongside possible computation speed gains.

References 1. Morris M. Spatial Models In Stan: Intrinsic Auto-Regressive Models for Areal Data 2019 [Available from: https://mc-stan.org/users/documentation/case-studies/icar_stan.html.

16: A Bayesian Analysis of FINEARTS-HF

Alasdair D Henderson¹, Brian L Claggett², Akshay S Desai², Mutthiah Vaduganathan², Carolyn S Lam³, Bertram Pitt⁴, Michele Senni⁵, Sanjiv J Shah⁶, Adriaan A Voors⁷, Faiez Zannad⁸, Meike Brinker⁹, Flaviana Amarante¹⁰, Katja Rohwedder¹¹, James Lay-Flurrie¹², Scott D Solomon², John JV McMurray¹, Pardeep S Jhund¹

¹BHF Glasgow Cardiovascular Research Center, School of Cardiovascular and Metabolic Health, University of Glasgow, Glasgow, Scotland, UK

²Cardiovascular Division, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA

- ³National Heart Centre Singapore & Duke-National University of Singapore, Singapore
⁴University of Michigan, School of Medicine, Ann Arbor, Michigan, USA
⁵University Bicocca Milan, Italy, Papa Giovanni XXIII Hospital, Bergamo, Italy
⁶Northwestern University Feinberg School of Medicine, Chicago, Illinois, USA
⁷University of Groningen, Groningen, Netherlands
⁸Université de Lorraine, Inserm Clinical Investigation Centre, CHU, Nancy, France
⁹Bayer AG, Research & Development, Pharmaceuticals, Wuppertal, Germany
¹⁰Cardiology and Nephrology Clinical Development, Bayer SA, São Paulo, Brazil
¹¹Global Medical Affairs, Berlin
¹²Bayer plc, Research & Development, Pharmaceuticals, Reading, UK

Background The FINEARTS-HF trial was a large, double-blind, placebo-controlled, randomised trial of the non-steroidal mineralocorticoid receptor agonist (MRA) finerenone. Conventional frequentist analysis of FINEARTS-HF found that finerenone reduced the primary composite endpoint of heart failure events and cardiovascular death in patients with heart failure with mildly reduced or preserved ejection fraction (HFmrEF/HFpEF) (rate ratio 0.84; 95% confidence interval, 0.74 to 0.95; $P = 0.007$). Bayesian methods offer alternative analytical approaches to provide probabilistic estimates of efficacy and safety, and flexibility to allow the inclusion of prior information and hierarchical modelling of subgroup effects. We analysed FINEARTS-HF with Bayesian methods to demonstrate the strengths and limitations compared to the primary frequentist analysis.

Methods In a pre-specified Bayesian analysis of FINEARTS-HF, we estimated treatment efficacy under a range of scenarios incorporating prior information from two trials of finerenone in participants with chronic kidney disease and type 2 diabetes (FIDELIO-DKD and FIGARO-DKD, pooled in the FIDELITY program) and a steroid MRA in patients with HFmrEF/HF-pEF (TOPCAT). We also used a combination of these trials in a robust meta-analytic prior. All models of the primary recurrent endpoint were analysed with Bayesian Cox proportional hazards models, with stratum-specific baseline hazards and hierarchical structure for subject-specific random effects. Secondary endpoints were analysed with Bayesian stratified Cox proportional hazards models. We used Bayesian hierarchical models to estimate subgroup effects with reduced heterogeneity from small sample sizes in frequentist subgroup analyses.

Results A total of 6,001 patients were included and the Bayesian analysis with vague priors confirmed the primary frequentist results with a 95% probability that the rate ratio was between 0.74 and 0.94. Including prior information from previous nonsteroidal and steroid MRA trials supported this finding and strengthened the probability of a beneficial treatment effect. Bayesian subgroup estimates were qualitatively similar to frequentist estimates but more precise and closer to the overall treatment effect. The probability that finerenone improves survival time until cardiovascular death was 79% (HR 0.93, 95% CrI: 0.79-1.09, $\text{Pr}(\text{HR}<1) = 79\%$), and all-cause mortality was 87% (HR 0.94, 95% CrI: 0.84-

1.05, $\text{Pr}(\text{HR}<1) = 87\%$), although any benefit was likely small on an absolute scale.

Conclusion The non-steroidal MRA finerenone reduced the rate of heart failure events and cardiovascular death, and there is a strong probability that there is a small reduction in CV death and all-cause mortality. Bayesian methods offer additional insights to the analysis of a large randomized control trial.

17: Fast Approximation of Joint Models: A Comparative Evaluation of Bayesian Methods

Jinghao Li, David M Hughes

University of Liverpool, United Kingdom

Background Joint models are widely employed in statistics to simultaneously analyze longitudinal data and time-to-event data, effectively capturing the dynamic relationships between the two processes. This framework has shown significant utility in biostatistics and clinical research. The widespread adoption of joint models enable clinicians to make predictions about patient specific risk that update over time, and aid clinical decision making. However, the increased complexity of joint models compared to separate longitudinal and survival models necessitates more sophisticated parameter estimation methods. Early contributions using Maximum Likelihood Estimation (MLE) laid the foundation for joint model estimation, followed by advancements in Bayesian methods that employed Markov Chain Monte Carlo (MCMC) techniques for inference. While MCMC-based approaches, such as JMBayes and rstanarm, provide accurate parameter estimates, they are computationally expensive and exhibit slow convergence, particularly when handling large datasets and multiple longitudinal variables. More recently, the INLAjoint package has been introduced, applying the Integrated Nested Laplace Approximation (INLA) to joint models, offering faster computation but with potential trade-offs in accuracy.

Method Variational Bayes (VB) inference, originally popularized in artificial intelligence applications, has gained increasing attention in statistical research due to its computational efficiency and scalability, as highlighted by Ormerod and Wand (2010). This study aims to provide a comprehensive evaluation of existing Variational Bayes methods for joint models, comparing their performance with established MCMC- and INLA-based approaches. The comparison focuses on key evaluation criteria, including computational efficiency, estimation accuracy, error rates, and convergence speed. Implementations from existing R packages, including Stan-based MCMC and Variational Bayes algorithms, are used in the analysis. Performance is assessed through simulation studies generated with the simsurv package (Brille-

man) under controlled conditions, as well as through validation on real-world data from the Primary Biliary Cirrhosis (PBC) study. **Results** The results will include a detailed comparison of model fitting times, estimation accuracy, error metrics, and convergence properties across the different approaches. The evaluation is ongoing, and comprehensive results will be presented at the conference. Future analyses will explore potential trade-offs in estimation bias and error, providing insights into the relative advantages of different inference methods for large-scale joint model applications. **Keywords** Joint Model, Variational Bayes, Bayesian Inference, MCMC, INLA, Longitudinal Data, Survival Analysis **References** Ormerod, J.T. and Wand, M.P. (2010). Explaining Variational Approximations. *The American Statistician*, 64(2), pp.140–153. doi: <https://doi.org/10.1198/tast.2010.09058>.

18: Confidence Intervals for Comparing Two Independent Folded Normals

Eleonora Di Carluccio¹, Sarah Ongutu², Ozkan Köse³, Henry G. Mwambi^{1,2}, Andreas Ziegler^{1,2,4,5}

¹Cardio-CARE, Medizincampus Davos, Davos, Switzerland

²School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Pietermaritzburg, South Africa

³Orthopedics and Traumatology Department, Antalya Training and Research Hospital, Antalya, Turkey

⁴Department of Cardiology, University Heart and Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

⁵Centre for Population Health Innovation (POINT), University Heart and Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

The absolute change in the angle measured immediately after surgery and after bone healing is a clinically relevant endpoint to judge the stability of an osteotomy. Assuming the difference in angles is normally distributed, the absolute difference follows a folded normal distribution. The confidence interval for the angle change of a novel fixation screw compared to a standard fixation screw may be used for evaluating non-inferiority. In this work, we suggest that the simple two-sample t-statistic or Welch-statistic may serve as the basis for confidence interval calculations for the difference between two folded normal. The coverage probabilities of derived confidence intervals are investigated by simulations. We illustrate the approaches with data from a randomized controlled trial and an observational study on hallux valgus i.e., bunion surgery. In the simulation studies, asymptotic and both non-parametric and parametric bootstrap confidence intervals based on the t-statistic and Welch-test were close to nominal levels. Methods based on the chi-squared distributions were not deemed appropriate for comparing two folded normal. We recommend using confidence intervals

based on the t-statistic or the Welch-statistic for evaluating non-inferiority in trials where the stability of angles after osteotomy is to be compared.

19: Towards Realistic Synthetic Nanopore Protein Signals: A Comparative Study of Stochastic and GAN-Based Methods

Göran Köber¹, Jonas Bürgel¹, Tobias Ensslen^{1,2}, Oliver Amft^{1,2}

¹University of Freiburg, Germany

²Hahn-Schickard, Germany

Nanopores provide a powerful tool for molecular analysis, enabling direct, single-molecule measurements of nucleotides, peptides, and other biopolymers. However, developing machine learning models for tasks like peptide sequence recognition is challenging due to the scarcity of labeled training data, as experimental data collection is both expensive and time-consuming. Synthetic data generation offers a promising solution by providing high-quality, customizable datasets for algorithm development and benchmarking.

We develop and compare several techniques for generating synthetic nanopore protein data, leveraging both stochastic methods and deep learning approaches, with a particular focus on Generative Adversarial Networks (GANs). The generated signals can be of arbitrary lengths, reaching up to hundreds of thousands of steps—far exceeding commonly reported time series lengths in the literature. The generation process is structured into two phases. First, a flat reference signal is synthesized to mimic the general shape of a blockade. Next, fluctuation generation algorithms introduce the fluctuating patterns of experimental data into the reference signal.

Multiple signal generation algorithms are explored, starting with a simple Gaussian noise model as a baseline. More advanced stochastic approaches, combining cubic interpolation with Gaussian noise, produce signals that closely resemble real blockade events. Additionally, an RNN-WGAN architecture is developed to generate arbitrarily long, high-fidelity signals that are challenging to distinguish from experimentally observed data. To evaluate the quality of generated signals, a discriminative score is computed using an RNN classifier, complemented by dimensionality reduction on established feature extraction libraries for time series data.

We also provide a comparative analysis of stochastic and data-driven methods, examining both their qualitative and quantitative differences and find that GAN-based methods achieve the best overall results. To the best of our knowledge, this work is the first to introduce high-quality synthetic nanopore protein sensing data generation methods, paving the way

for advanced machine learning applications and addressing the critical need for labeled, customizable synthetic datasets in the field.

20: Data Transformations in Machine Learning Approaches for Studying Microbiota as a Biomarker of Non-Response Risk to CFTR Modulators

Marta Avalos¹, Céline Hosteins¹, Diego Kauer¹, Chloé Renault¹, Raphaël Enaud², Laurence Delhaes²

¹University of Bordeaux - Inria - Inserm BPH U1219, France

²University of Bordeaux, CHU Bordeaux, Inserm U1045, France

Cystic fibrosis (CF) is a genetic disease caused by mutations in the CF transmembrane conductance regulator (CFTR) gene. Impaired mucociliary clearance and the accumulation of respiratory secretions, combined with an altered immune response and chronic treatments, disrupt the airway microbiota and mycobiota. These dysbioses, characterized by reduced microbial diversity and a predominance of opportunistic pathogens, correlate with disease severity and may serve as biomarkers for disease progression.

The introduction of CFTR modulator therapies has transformed CF management, significantly altering the disease's clinical course by enhancing mucosal hydration and improving patient outcomes. However, response to these therapies remains highly variable among patients, underscoring the need for predictive biomarkers. The airway and digestive microbiota, which play a crucial role in disease progression, represent promising candidates. While bacterial and fungal dysbioses in CF are well documented, their potential as biomarkers for predicting therapeutic response remains poorly explored, posing significant methodological challenges.

Microbiome studies in CF typically involve small cohorts and high-dimensional data, often compositional, zero-inflated, and sometimes longitudinal. Moreover, integrating heterogeneous data sources—including bacterial and fungal communities from different anatomical sites (lung and gut) alongside clinical factors—is essential for building robust predictive models. This requires advanced statistical and machine learning approaches to address challenges in feature selection, model interpretability, and data integration.

In this study, based on CF patients from the French LumivaBiota cohort, we examine how transformations of relative abundance data affect both the performance and interpretability of various linear (Lasso, PLS, PCA regression) and non-linear (SVM, Random Forest, Neural Networks) machine learning methods. We compare these approaches in their ability to predict

non-response to CFTR modulators, balancing the trade-off between model complexity and interpretability—a key consideration for clinical application.

Our findings provide insights into best practices for microbiome-based predictive modeling in CF and offer methodological guidance on selecting appropriate data transformations and machine learning frameworks for biomarker discovery in high-dimensional biological datasets.

21: On Microbiome Data Analysis using Bayesian Method under the Assumption of a Zero-Inflated Model

Yuki Ando, Asanao Shimokawa

Tokyo University of Science, Japan

Background / Introduction

The data on the abundance of microbial groups is called microbiome data. One of the purposes of analysing microbiome data is to compare the abundance of microbes in the bodies of test subjects with different conditions. There are two main characteristics of microbiome data: firstly, the abundance is discrete, and secondly, there are an excessive number of zeros in the data. However, in order to make it possible to compare between test subjects with different total abundances of microbes, the abundance is sometimes converted into a proportion. This is the method we will use in this study. In this case, the abundance takes a continuous value in the range from 0 to 1. The zero-inflated beta model is the most commonly used population distribution for abundance. This is a distribution in which the abundance follows a beta distribution with a certain probability, and takes 0 in other cases. Furthermore, Chen and Li (2016) stated that the probability that the abundance follows a beta distribution and the parameters of the beta distribution are expressed by a logistic regression model with the covariates of the subjects as explanatory variables. We will examine the method of estimating the parameters in this model.

Methods

The parameters of the zero-inflated beta model are currently estimated using the maximum likelihood method. However, since it is not possible to obtain the maximum likelihood estimate analytically, we will use an iterative calculation method such as the EM algorithm in combination. One drawback of this method is that it cannot obtain good estimates when the sample size of the microbiome data is small, i.e., when the number of subjects is small. Therefore, we consider estimating the parameters using Bayesian methods.

Results

We applied the maximum likelihood method and Bayesian methods to simulation data and compared the obtained estimates. We found that the Bayesian method worked well in situations with small sample sizes.

Conclusion

We dealt with parameter estimation when assuming a zero-inflated beta model for microbiome data.

We found that it is recommended to use the Bayesian method rather than the maximum likelihood method for microbiome data with a small sample size.

Reference

Chen E.Z. and Li H. 2016. A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics* 32 (17): 2611–2617.

22: Estimating Prevalence of Cystic Fibrosis Heterozygosity using Fast Boltzmann Inversion (FBI): An Improved Monte Carlo Algorithm for Bayesian Inference

Jan Brink Valentin

Danish Center for Health Services Research, Department of Clinical Medicine, Aalborg University, Denmark

Background Monte Carlo sampling of probabilistic potentials is often biased if the density of state is not properly managed. Bias is further induced when applying shrinkage priors to avoid overfitting. Boltzmann inversion (BI) provides a generic sampling scheme to avoid such bias. However, because of the iterative nature of BI, the algorithm is often intractable. In this study, we developed a fast Boltzmann inversion (FBI) algorithm with the same computational complexity as the standard Metropolis-Hastings (MH) algorithm and applied the method for estimating the heterozygous carrier prevalence of Cystic fibrosis (CF).

Case: CF is a rare genetic disease which have a massive impact on the patients' health, daily living and overall survival. The disease is inherited from both parents, and in Denmark children have been screened for CF at birth since 2016. While the incidence rate can be estimated using patient registers, the heterozygous carrier prevalence is not easily found.

Method We applied a simple two parameter probabilistic model for the probability of having CF conditioning on being the first-born child in a family. This probability was considered the target distribution, and the model parameters included the proportion of heterozygous

carriers. We linked the Danish national patient register with the central person register to estimate the mean and variance of the target distribution. We applied shrinkage priors for the model parameters with low, moderate, and strong shrinkage, to avoid overfitting. The FBI algorithm was then used to estimate the model parameters, and the results were compared to that of the MH algorithm.

Results Using the register data the target probability was estimated to be 2.38 per 10.000. The MH algorithm with low, moderate and strong shrinkage were biased and missed the target by 0.16, 0.25 and 0.35 per 10.000. The FBI algorithm with low and moderate shrinkage were both on target with less than 0.001 bias and estimated the proportion of heterozygous carriers in the Danish population to be 3.05 percent ($SE = 0.54$). However, the FBI algorithm with strong shrinkage did not converge. Finally, the FBI algorithm used the same computational time as the MH algorithm.

Conclusion The FBI algorithm provides an unbiased estimate when applying shrinkage estimators without increasing computational time compared to other Monte Carlo algorithms. In addition, the FBI algorithm reduces the issue of setting the hyper parameters of the prior distributions in a Bayesian context.

23: A Sparse Graph Representation of Hi-C Data for Colorectal Cancer Prediction

Jiwon Im¹, Mingyu Go², Insu Jang³, Minsu Park¹

¹Department of Statistics and Data Science, Chungnam National University, Republic of Korea

²Graduate School of Data Science, KAIST, Republic of Korea

³Korea Research Institute of Bioscience and Biotechnology, Republic of Korea

Colorectal cancer (CRC) remains a leading cause of cancer-related morbidity and mortality, emphasizing the critical need for early and precise predictive modeling. While advances in genomics and deep learning have enabled computational cancer classification, existing models often face challenges in capturing the complexity of chromatin organization and high-dimensional genomic data.

This study presents a graph-based predictive framework utilizing high-throughput chromosome conformation capture (Hi-C) data from chromosome 18, a region implicated in CRC pathogenesis. The method constructs a sparse weighted graph from chromatin interactions and applies a graph neural network for classification. An optimal bandwidth selection technique removes redundant connections while retaining key genomic relationships to enhance

computational efficiency and interpretability.

Experimental evaluations on real-world Hi-C datasets indicate that the proposed approach achieves competitive classification accuracy while improving F1-score and precision-recall performance with reduced training complexity. These findings suggest that sparse graph-based Hi-C analysis may be a useful framework for CRC prediction and contribute to graph representation learning in genomic medicine.

Keywords Hi-C, graph neural network, sparse graph representation, CRC classification

24: A Multi-State Survival Model for Recurring Adenomas in Lynch Syndrome Individuals

Vanessa García López-Mingo, Veerle Coupé, Marjolein Greuter, Thomas Klausch

Amsterdam UMC, Netherlands, The

Introduction Lynch syndrome is a genetic condition that predisposes individuals to develop colorectal cancer (CRC). It is characterized by a deficiency in the mismatch repair (MMR) system occurring early in life, leading to increased risk of accumulating DNA damage. Individuals with Lynch develop adenomas, pre-cursor lesions to CRC, in the bowel at a higher rate compared to general population. This necessitates close surveillance of affected individuals by colonoscopy. Although surveillance intervals are short (one to three years), continued surveillance is needed to manage CRC risk throughout life.

Based on surveillance data from Lynch patients, this study aims to estimate the time to repeated non-advanced adenoma (nA) formation and progression to advanced adenomas (AA) or CRC. We develop a novel multi-state survival model that, contrary to available models, handles recurring adenomas that characterize Lynch.

Methods The model treats the adenoma formation as panel count data, where the occurrence of recurrent adenomas is observed only at a sequence of discrete time intervals (colonoscopies). Specifically, the development of nA is modelled as a Poisson process, but with a modification to account for the delay associated with the occurrence of MMR deficiency around the time of the first nA, incorporated through a Weibull model. Immediately after, a Poisson process for later nAs is initialized. Furthermore, every adenoma is assumed to progress to AA or CRC, where the sojourn time is also modelled Weibull distributed. All sojourn times are regressed on covariates like sex and the affected gene to uncover heterogeneity. A Bayesian Metropolis-within-Gibbs sampler combined with data augmentation for

the latent times is employed to estimate the parameters.

Results In first Monte Carlo simulations, we found good performance of the estimation procedure, with unbiased estimates and good mixing across the chains. Additionally, the coverage percentages of the credible intervals matched the nominal level of 95%. At ISCB we will present details on the application to the Lynch patient data, which is currently under development.

Conclusion This study presents a novel model for analysing adenoma development in Lynch accounting for the impact of MMR deficiency. By using the combination of the delay on the first adenoma and a Poisson process for the recurrent ones we capture the dynamics of adenoma development in Lynch more accurately than existing multi-state screening models such as “msm” (Jackson, 2011). Developing dedicated models for disorders like Lynch could help improve prevention of CRC in affected groups.

25: Evaluating Completeness of Data in CPRD’s Breast Cancer Data: Implications for External Controls for Surrogate Endpoints

Dorcas N Kareithi^{1,3}, Jingky Lozano-Kuehne¹, David Sinclair², James Wason¹

¹Biostatistics Research Group, Newcastle University, United Kingdom

²Older People and Frailty Policy Research Unit, Newcastle University, United Kingdom

³Jasiri Cancer Research Foundation Kenya

Background Registries such as CPRD Aurum and national cancer registries offer valuable sources of observational data that can be used as external or historical controls for cancer clinical trials and other health research. However, an extensive review of evidence has shown that the investigation of alternative measurements from routinely collected data is dependent on access, validity, and completeness of such data. This study evaluates the completeness, patterns and impact of missing data in breast cancer patients using data from CPRD Aurum and Cancer Registration and Treatment datasets.

Methods We used linked datasets from CPRD Aurum, the Tumor Registration dataset, and the Treatment Characteristics dataset to identify and extract breast cancer cases (ICD-10: C50). Key clinical variables including demographic characteristics, tumour type and size, comorbidity score; tumour screening, tumour treatment, and cancer stage from female patients who were a18 and above in 2005, were analysed. Completeness of data in 6months' follow up periods post diagnosis from 2005-2024 and patterns of missing data were assessed using descriptive statistics and Little's MCAR test to determine missingness mechanisms. No

imputation methods were applied, as the focus was on understanding completeness and the extent or impact of missingness.

Results Preliminary findings of 2.9M records from 68,613 participants who fit our inclusion and exclusion criteria indicate high completeness (>90%) in most demographic characteristics, most observation event dates except hospitalisation date, high (>90%) completeness in most tumour stage and characteristics except for PR and ER scores, high (>90%) completeness in most tumour treatment variables, and average (>60%) completeness of quality-of-life variables. Preliminary time-to-event analyses suggest that incomplete data used to compute surrogate outcomes, such as the quality-of-life data could affect the derivation, computation and estimation of key established surrogate endpoints such as Disease-Free survival (DFS), Time to Next Treatment (TTNT), Event Free Survival (EFS) and Overall survival.

Conclusion The preliminary findings highlight the value of registry data for use as external or historical controls for cancer clinical trials and other health research, but caution against potential biases introduced by some of the incomplete data, which may impact clinical interpretations and policy decisions.

26: Identification of Risk Factors for the Development of De Novo Malignancies after Liver Transplantation

Tereza Hakova¹, Pavel Taimr¹, Tomáš Hucl¹, Zdeněk Valenta²

¹Dept. of Hepatogastroenterology, Institute of Clinical and Experimental Medicine, Prague, Czechia

²Dept. of Statistical Modelling, Institute of Computer Science of the Czech Academy of Sciences, Prague, Czechia

Background De novo malignancies (DNM) are a significant long-term complication in liver transplantation, immunosuppressive therapy being a key contributing factor. While necessary to prevent graft rejection, exposure to immunosuppressants may increase the risk of post-transplant malignancies. Identifying risk factors for DNM is crucial to improving post-transplant management strategies. This study uses the cohort of liver transplant patients aged 18 years and older to study competing risks of the incidence of DNM or death, focusing on the role of cumulative exposure to immunosuppressants. Independent prognostic factors, such as the age of donor/recipient, gender, smoking, diabetes status, etc. were adjusted for in the models. We hypothesised that high cumulative doses of immunosuppressants could correlate with an increased incidence of malignancies, suggesting the need for individualised immunosuppression strategies.

Methods Retrospective right-censored and left-truncated cohort data on 1,073 liver transplant patients aged 18 years or more were used to study competing risks of the incidence of cancerous disorders and death in the context of immunosuppression following transplantation (TX). We studied the effect of cumulative exposure to several immunosuppressants (Azathioprin, Cyclosporin A, Mycophenolate Mofetil, Prednison, Simulect, Sirolimus and Tacrolimus), adjusted for possible confounders. Cause-specific survival models, incl. the Cox PH model, Aalen's additive model and its McKeague-Sasieni extension, were employed in analysing the effect of left-truncated time-dependent immunosuppression doses on the right-censored outcomes. Follow-up period was limited to 10 years.

Results Results for time-to-malignancy data showed that male gender and higher recipient and donor age were associated with elevated hazard of malignancy incidence. Immunosuppression using Mycophenolate, Sirolimus and Tacrolimus was associated with the reduction in the hazard, while Simulect had an adverse effect on malignity incidence.

The age, smoking and male gender of TX recipient had an adverse effect on the hazard of death. Applications of Mycophenolate, Prednison, Sirolimus and Tacrolimus all proved to maintain a protective effect on the incidence of death. The latter drug had a time-varying protective effect, the strongest being for a few months after TX.

Conclusion Our results bring new insights into immunosuppressive treatment. Inconsistency with published studies may be due to a different methodology and the patient population. This study highlights the importance of monitoring immunosuppressive drug levels and controlling modifiable risk factors, such as smoking, in liver transplant recipients. Understanding the multifactorial nature of post-transplant malignancies can lead to improved patient management.

27: Are ACE Inhibitors Associated with Increased Lung Cancer Risk, or are Unmeasured Confounders Biasing Results?

Sean Maguire, Ruth Keogh, Elizabeth Williamson, John Tazare

London School of Hygiene & Tropical Medicine, United Kingdom

Background Unmeasured confounding is nearly always a concern in observational studies of treatment effects. However, despite methods being available to assess the potential impact, it is often ignored. We illustrate methods for assessing the impact of unmeasured confounding through a study of ACE inhibitors and ARBs; commonly prescribed drugs for the treatment of high blood pressure. Safety concerns for ACE inhibitors raised by observa-

tional study findings of higher lung cancer risks in ACE inhibitor users relative to ARB users, and the inconsistent findings in subsequent observational studies, may have been caused by unmeasured confounding.

Methods Using data from the Clinical Practice Research Datalink, we identified a cohort of UK adults who initiated an ACE inhibitor or ARB treatment for the first time between 1995 – 2019, and fitted a Cox model for the outcome of lung cancer, with adjustment for a number of measured confounders. A conditional hazard ratio was estimated, accounting for competing events.

E-values were used to quantify the potential impact of unmeasured confounding on the effect estimates. E-values, introduced in 2017 by VanderWeele and Ding, quantify the minimum strength of association an unmeasured confounder (or set of unmeasured confounders) would need to have with ACE inhibitor use and lung cancer incidence, in order to ‘tip’ our results and change our conclusions.

Covariate e-values, introduced by D'Agostino McGowan and Greevy in 2020, were also calculated and contextualise the potential impact of unmeasured confounding in previous ACE inhibitor and lung cancer studies in the literature.

Results Our cohort contained 984,000 initiators of ACE inhibitor/ARB users. We found no evidence that ACE inhibitor use is associated with increasing lung cancer risk (conditional hazard ratio = 0.997, 95% CI 0.924 – 1.076). Our investigations using e-values show that this result could be easily tipped to a significantly harmful or protective effect. Similar results are found for the previous studies in the literature.

Conclusion Through use of quantitative bias analyses using e-values we found that it is likely that studies which have reported both protective and harmful effects of ACE inhibitors, are biased due to unmeasured confounding.

28: DO GENETIC CHANGES in 15q13.3 MEAN LOWER IQ SCORE?

Tadas Žvirblis¹, Pilar Caro², Audronė Jakaitienė¹, Christian Schaaf²

¹Institute of Data Science and Digital Technologies, Vilnius University

²Institute of Human Genetics, Heidelberg University

Background Genetic changes affecting the copy number of chromosome 15q13.3 have been associated with a group of rare neurodevelopmental conditions (autism spectrum disorder, epilepsy, schizophrenia, and others) [1]. The critical region contains approximately 10 genes. Treatments are limited and are restricted to targeting the main symptoms rather than the

underlying etiology. Not every person harboring a 15q13.3 copy number change will manifest the disease, and the severity and clinical diagnosis are difficult to predict [2]. This represents a significant challenge in modelling and determining health outcomes.

Methods Multi-center prospective study was conducted to assess cerebral activity and neural network function alterations in individuals with 15q13.3 microdeletion or microduplication. It was planned to enroll 15 subjects for each aberration, as well as 15 healthy subjects. During the study period electrophysiological brain network analysis, IQ testing and detailed genetic analysis were performed for each subject. All subjects provided written informed consent. Study protocol was approved by the Ethics Board of the Medical Faculty of Heidelberg University No. S-212-2023.

Results Six subjects with genetic changes affecting the copy number of chromosome 15q13.3 were identified during the interim statistical analysis. Five (83.3%) had a deletion of 15q13.3 and 1 (16.7%) duplication. The mean (SD) age was 27.5 (11.36) years; 2 (33.3%) of them were of adolescent age, and half (50.0%) of subjects were male. The mean (SD) IQ score was 76.7 (18.14), and it was statistically significantly lower than the average population IQ score ($p = 0.027$). The mean (SD) IQ score for the males was slightly higher than that of females: 79.7 (22.03) vs. 73.7 (17.62) for males and females, respectively.

Conclusion The interim statistical analysis showed that subjects with 15q13.3 microdeletion or 15q13.3 microduplication have lower IQ score than average population.

Funding. This work is part of the EJP RD project “Resolving complex outcomes in 15q13.3 copy number variants using emerging diagnostic and biomarker tools (Resolve 15q13)” No. DLR 01GM2307 and has received funding from EJP RD partner the Research Council of Lithuania (LMTLT) under grant agreement No. S-EJPRD-23-1.

Keywords rare neurodevelopmental conditions, 15q13.3 microdeletion, 15q13.3 microduplication

References [1] Gillentine MA, Schaaf CP. The human clinical phenotypes of altered CHRNA7 copy number. Biochem Pharmacol. 2015;97(4):352-62.

[2] Yin J, Chen W, Yang H, Xue M, Schaaf CP. Chrna7 deficient mice manifest no consistent neuropsychiatric and behavioral phenotypes. Sci Rep. 2017 Jan 3;7:39941. PMCID: PMC5206704

29: Measuring the Performance of Survival Models to Personalize Treatment Choices

Orestis Efthimiou¹, Jeroen Hoogland², Thomas Debray³, Valerie Aponte Ribero¹, Wilma Knol⁴, Huiberdina Koek⁴, Matthias Schwenkglenks⁵, Séverine Henrard⁶, Matthias Egger⁷, Nicolas Rodondi¹, Ian White⁸

¹University of Bern (Switzerland)

²Amsterdam University Medical Centers (The Netherlands)

³Smart Data Analysis and Statistics B.V. (The Netherlands)

⁴Utrecht University (The Netherlands)

⁵University of Basel (Switzerland)

⁶UCLouvain (Belgium)

⁷University of Bern (Switzerland), University of Bristol (UK), University of Cape Town (South Africa)

⁸University College London (UK)

Background Statistical and machine learning algorithms can be used to predict treatment effects at the participant level using data from randomized clinical trials (RCTs). Such predictions can facilitate individualized treatment decisions. Although various methods have been proposed to assess the accuracy of participant-level treatment effect predictions, it remains unclear how they can be applied to survival data.

Methods We propose new methods to quantify individualized treatment effects for survival (time-to-event) outcomes. First, we describe alternative definitions of participant-level treatment effects for survival outcomes. Next, we summarize existing and introduce new measures to evaluate the performance of models predicting participant-level treatment effects. We explore metrics for assessing discrimination, calibration, and decision accuracy of such predictions. These generic metrics are applicable to both statistical and machine learning models and can be used during model development (e.g., for model selection or internal validation) or when testing models in new settings (e.g., external validation). We illustrate our methods using both simulated data as well as real data from the OPERAM trial, an RCT involving multimorbid older adults randomized to either standard care or a pharmacotherapy optimization intervention. We fit competing statistical and machine learning models and apply our newly developed methods to compare their performance.

Results Analyses of simulated data demonstrated the utility of our metrics in evaluating the performance of models predicting participant-level treatment effects. Application in OPERAM revealed that the models we developed performed sub-optimally, with moderate-to-poor performance in calibration and poor performance in discrimination and decision accuracy, when predicting individualized treatment effects.

Conclusion Our methods are applicable for models aimed at predicting participant-level treatment effects for survival outcomes. They are suitable for both statistical and machine learning models and can guide model development, validation, and potential impact on decision making.

30: A Framework for Estimating Quality Adjusted Life Years using Joint Models of Longitudinal and Survival Data

Michael Crowther¹, Alessandro Gasparini¹, Sara Ekberg¹, Federico Felizzi², Elaine Gallagher³, Noman Paracha³

¹Red Door Analytics AB, Stockholm, Sweden

²Department of Computer Science, ETH Zurich, Switzerland

³Bayer Pharmaceuticals, Basel, Switzerland

Background Quality of life (QoL) scores are integral in cost-effectiveness analysis, providing a direct quantification of how much time patients spend at different severity levels. There are a variety of statistical challenges with modeling and utilizing QoL data appropriately. QoL data, and other repeatedly measured outcomes such as prostate-specific antigen (PSA), are often treated as time-varying covariates, which only change value when a new measurement is taken - this is biologically implausible. Additionally, such data often exhibits both between and within subject correlations, which must be taken into account, and are associated with survival endpoints. The proposed framework utilizes "progression" or similar intermediate endpoints or biomarkers like EQ-5D, and models them jointly with overall survival, allowing us to directly calculate quality adjusted life years (QALYs).

Methods Motivated by the prostate cancer trial setting, we simulated data representing repeatedly measured PSA levels, utilities and overall survival. Using numerical integration and the delta method, we then derive analytical estimates of QALYs, differences in QALYs and restricted time horizon QALYs from the estimated multivariate joint model, along with uncertainty.

Results PSA and utilities were modeled flexibly using linear mixed effects submodels with restricted cubic splines to capture the nonlinear development over follow-up time. An interaction with treatment was also included to allow different trajectories in those treated and those on placebo. Both PSA and utility were linked to survival through their current value and slopes, with a Weibull survival submodel. Treatment was estimated to provide an additional 1.074 QALYs (95% CI: 0.635, 1.513) across a lifetime horizon.

Conclusion Deriving QALYs from a joint model of longitudinal and survival data accounts for all of the statistical and biological intricacies of the data, providing a more appropriate, and accurate, estimate for use in cost-effectiveness modeling, and hence reducing uncertainty.

31: Cut Off Determination using Model Derived Estimate in Survival Prediction Model

Jungbok Lee

Asan Medical Center & Univ of Ulsan College of Medicine, South Korea

For practical clinical applications, various prediction models related to disease onset, risk, prognosis, and survival are being developed using EMR or clinical research data. The score generated by these models is used as a measure of risk, often categorized for practical purposes. Determining an appropriate cutoff for score categorization has become a topic of interest. For example, in the case of time-to-event outcomes, the most intuitive method is to identify a cutoff that maximizes the log-rank test statistic. However, the method based on test statistics has the limitation that the cutoff may vary depending on the distribution of the dataset used for model building.

In this study, we present:

- 1) A phenomenon where the cutoff varies when there are relatively many or few high- or low-risk subjects in the training set.
- 2) A hazard estimation procedure using a piecewise hazard model and resampling method for survival data.
- 3) Cutoff criteria for when the hazard rate estimated by the model follows linear, parabolic, cubic, logarithmic, or exponential curves.
- 4) A proposed resampling procedure to account for variation in the distribution of events, based on the initially estimated cutoff value.

Determining cutoffs based on tests in survival data is dependent on the distribution of scores, censoring rates, and sample size. The method using model-derived estimates can help adjust for these dependencies. The term "optimal" in cutoff determination is limited to the original dataset and the testing method used to identify the cutoff.

32: A Two-Step Testing Approach for Comparing Time-to-Event Data under Non-Proportional Hazards

Jonas Brugger^{1,2}, Tim Friede², Florian Klinglmüller³, Martin Posch¹, Franz König¹

¹Medical University of Vienna, Vienna, Austria

²University Medical Center Göttingen, Germany

³Austrian Agency for Health and Food Safety, Vienna, Austria

The log-rank test and the Cox proportional hazards model are commonly used to compare time-to-event data in clinical trials, as they are most powerful under proportional hazards. But there is a loss of power if this assumption is violated, which is the case for some new oncology drugs like immunotherapies. We consider a two-stage test procedure, in which the weighting of the log-rank test statistic depends on a pre-test of the proportional hazards assumption. I.e., depending on the pre-test either the log-rank or an alternative test is used to compare the survival probabilities. We show that if naively implemented this can lead to a substantial inflation of the type-I error rate. To address this, we embed the two-stage test in a permutation

test framework to keep the nominal level alpha. We compare the operating characteristics of the two-stage test with the log-rank test and other tests by clinical trial simulations.

33: A Systematic Review of Bayesian Survival Modelling for Extrapolating Survival in Cost-Effectiveness Analysis

Farah Erdogan¹, Gwénaël Le Teuff^{1,2}

¹Oncostat, INSERM U1018, France

²Department of Biostatistics and Epidemiology, Gustave Roussy, Paris-Saclay University, France

Background Cost-effectiveness analysis (CEA) aims to evaluate the clinical and economic impact of health interventions. In some settings, such as oncology, CEA requires estimation of long-term benefits in terms of life years. Survival extrapolation is then necessary when clinical trials have limited follow-up data. As highlighted in the NICE Technical Support Document (TSD) 21 (2020), Bayesian approach offers a flexible framework for incorporating external information in survival modelling and addressing uncertainty in survival prediction. This work aims to report how Bayesian is used to incorporate external information for extrapolating long-term survival in CEA. Methods: We conducted a systematic review up to

October 2024 to identify both methodological and non-methodological studies using different electronic databases (PubMed, Scopus, ISPOR conference database), completed by handsearching references cited in the automatically identified studies. Results: Of 52 selected studies (77% automatically and 23% manually), 52% were published since 2022 and 90% ($n=47$) focused on oncology. 52% ($n=27$) represented articles and 38% ($n=20$) were methodological works. 87% ($n=45$) used external data from different sources: clinical, registry, epidemiology, real world data, general population mortality and 17% ($n=9$) experts elicitation. We classified the studies into four non-mutually exclusive categories of Bayesian modelling (C1-C4). The first three categories combine, in order of increasing complexity, survival modelling and Bayesian formulation for incorporating external information. C1 (27%, $n=14$) includes standard parametric models (SMPs: exponential, Weibull, Gompertz, log-normal, log-logistic, generalized gamma distributions) with prior of parameters informed by historical data. C2 (48%, $n=25$) includes (i) a Bayesian multiple parameter evidence synthesis that allows to combine trial and external data, (ii) joint modelling of progression-free survival and overall survival, and (iii) non-SPMs (e.g., mixture and cure models). C3 (27%, $n=14$) groups complex hazard regression models (e.g., poly-hazard, relative survival) incorporating disease-specific and general population mortality with predominantly non-informative priors distributions on parameters. The last category (C4, 13%, $n=7$) represents Bayesian model averaging that weights predictions of different survival models by posterior model probabilities to address structural uncertainty. Conclusion: This review highlights the broad spectrum of Bayesian survival models and the different ways to incorporate external information, resulting in reduced uncertainty in survival extrapolation. Future research should focus on comparing these methods to identify the most suitable approaches given the intervention mechanisms and external data availability. This will help to standardize the use of Bayesian statistics for survival extrapolation and provide guidance, as proposed in the NICE TSD 14 on survival model selection procedures.

34: Power Comparison of Hazard Ratio Versus Restricted Mean Survival Time in the Presence of Cure

Ronald Bertus Geskus^{1,2}

¹Oxford University Clinical Research Unit, Vietnam

²University of Oxford, United Kingdom

Background The log-rank test and the Cox proportional hazards model lose power with non-proportional or crossing hazards. A large simulation study did not show consistent superior performance of restricted mean survival time (RMST) over the log-rank test in such settings [2]. That study did not consider the presence of cure, nor the presence of an independent

predictor of survival.

In a randomized controlled trial (RCT) investigating the effect of dexamethasone on survival in patients with tuberculous meningitis (TBM), the hazard ratio was the primary effect measure [1]. Baseline MRC grade strongly affected 12-month survival. Testing for difference in RMST gave lower p-value than the hazard ratio: 0.14 versus 0.22, and 0.075 versus 0.21 when correcting for MRC grade. We performed a simulation study to investigate gain in power of RMST.

Methods For each scenario we simulated 3000 data sets from Weibull distributions with two treatment arms and a three-level categorical predictor of survival representing MRC grade. Weibull parameters were estimated based on the RCT, after exclusion of the 12-month survivors. Sample size including survivors was 700. We also considered approximate scenarios assuming proportional hazards in the non-survivors; note that proportionality is lost once the survivors are included. Models with and without interaction with MRC grade were fitted. In an additional scenario we generated data with divergent survival curves, then converging at 12 months, and mortality between 50% and 100%.

Results All numbers refer to power, computed as the percentage of simulation runs that gave p-value below 0.05 for the test of treatment effect. With parameters according to the TBM data set, RMST outperforms the hazard ratio (43% versus 36%). Further improvement is seen with adjustment for MRC grade (51% versus 33%). Similar results are observed with data generated assuming proportional hazards for the non-survivors. The test for non-proportionality has power between 10% and 30%. In the additional scenario with 50% mortality, proportional hazards had much lower power than RMST (36% versus 98%), while power was similar with 100% mortality.

Conclusion Relative performance of proportional hazards versus RMST strongly depends on the shape of the survival curve and the presence of cure.

References

- [1] Donovan et al., Adjunctive dexamethasone for tuberculous meningitis in hiv-positive adults, New England Journal of Medicine 389 (2023), 1357-1367.
- [2] Dormuth et al., A comparative study to alternatives to the log-rank test, Contemporary Clinical Trials 128 (2023).

35: Sample Size Calculation in Prognostic Studies: A Comparative Analysis

Gloria Brigiari, Ester Rosa, Giulia Lorenzoni, Dario Gregori

Unit of Biostatistics, Epidemiology and Public Health, University of Padova, Italy

Introduction

In classical survival analysis, sample size estimation is typically based on risk differences or hazard ratios (HR) between patient groups. While widely used, these methods have limitations, such as assuming group independence, proportional hazards, and neglecting side-variables. To address these challenges, alternative approaches, such as those proposed by Riley et al. (2019, 2021), focus on model precision and the inclusion of covariates. However, there is no guarantee that these methods will lead to a singular conclusion on the required sample size. This study aims to evaluate the performance of traditional HR-based methods and Riley's precision-focused approach through sensitivity analysis and Monte Carlo simulations. The goal is to identify the ideal sample size and assess how the inclusion of covariates impacts model performance.

Methods

We conducted a sensitivity analysis using Monte Carlo simulations to compare classical HR-based sample size estimation methods with Riley's model precision approach. Simulations were run based on historical data, focusing on proportional hazards and the inclusion of multiple covariates. Once the appropriate sample size was determined, Riley's methodology was applied to evaluate the number of predictors that could be included in the model without overfitting. The analysis used a shrinkage factor of 0.9 to balance model complexity and accuracy. Finally, with the aim of assessing whether the calculated sample size allows for the generalizability of a previously developed model, a simulation-based method was applied to estimate the achieved precision, in terms of calibration, based on the given sample size.

Results

Traditional methods struggled to capture model complexity and did not consider relevant covariates effectively. In contrast, Riley's method allowed for the inclusion of more covariates while maintaining statistical robustness. The application of Riley's methodology revealed that the number of predictors that could be included without overfitting depended on the desired model accuracy metrics. External validation approach confirmed the adequacy of the calculated sample size, achieving good calibration and predictive accuracy of the model.

Conclusion

This study highlights the limitations of traditional HR-based methods and demonstrates the advantages of the proposed approach, which prioritizes model precision and avoids overfitting. By allowing the inclusion of additional covariates without sacrificing power, this methodology offers a flexible and reliable framework for sample size estimation and model development in prognostic studies.

36: Prognostic Score Adjustment in a Two-Slope Mixed Effects Model to Estimate Treatment Effects on eGFR Slope in CKD Patients

Silke Janitz¹, Maike Ahrens², Sebastian Voss², Bohdana Ratitch³, Nicole Rethemeier⁴, Meike Brinker⁴, Paula Vesterinen⁵, Antigoni Elefsinioti¹

¹Bayer AG, Germany

²Chrestos GmbH, Essen, Germany

³Bayer Inc., Mississauga, Ontario, Canada

⁴Bayer AG, Wuppertal, Germany

⁵Bayer AG, Espoo, Finland

Background The CHMP recently recognized the estimated glomerular filtration rate (eGFR) slope as a validated surrogate endpoint for clinical trials of treatments for chronic kidney disease (CKD). A common method for analysis of this endpoint is a two-slope linear spline mixed effects model (Vonesh et al., 2019). This model can serve as the primary analysis in future CKD trials with the option to adjust for baseline covariates, e.g., sodium-glucose cotransporter-2 inhibitor (SGLT2i) use and urinary albumin-to-creatinine ratio (UACR). Following a CHMP Qualification Opinion on prognostic covariate adjustment, we explore the potential benefits of integrating a prognostic score in the two-slope model using a historical database from two large CKD phase III studies FIDELIO-DKD and FIGARO-DKD.

Methods Using the FIGARO-DKD study, we developed prognostic score models via random forest methodology, focusing on patients receiving placebo. These models included approximately 60 baseline covariates. We conducted extensive simulations based on FIDELIO-DKD to assess potential precision gains in treatment effect estimates from including a prognostic score obtained for each participant as a prediction from an aforementioned prognostic model.

Results Pseudo simulations from FIDELIO-DKD indicated that integrating the prognostic score into a two-slope model without other covariates yielded moderate precision gains. When compared to a model, which included SGLT2i use and UACR category, the additional precision gains from including the prognostic score were reduced.

Conclusion While prognostic score adjustment can enhance efficiency in clinical trials, it has primarily been studied within classical linear models. This work explores prognostic score adjustment to a more complex model, illustrating how sponsors can utilize historical data for pseudo simulations to evaluate the utility of prognostic score adjustments in future trials. Based on our historical studies, our findings from pseudo simulations suggest that incorporating a prognostic score in addition to other key baseline covariates (such as SGLT2i use and UACR category) may not yield substantial additional efficiency in estimating treatment

effects.

Literature

Vonesh E, et al. Mixed-effects models for slope-based endpoints in clinical trials of chronic kidney disease. *Stat Med*. 2019;38(22):4218-4239.

European Medicines Agency. Qualification opinion for Prognostic Covariate Adjustment (PROCOVATM). Committee for Medicinal Products for Human Use (CHMP). 2022.

European Medicines Agency. Qualification opinion for GFR Slope as a Validated Surrogate Endpoint for RCT in CKD. Committee for Medicinal Products for Human Use (CHMP). 2023.

37: Comparative Effectiveness of ACE Inhibitors and Angiotensin Receptor Blockers to Prevent or Delay Dementia: a Target Trial Emulation

Marie-Laure Charpignon¹, Max Sunog², Colin Magdamo², Bella Vakulenko-Lagun³, Ioanna Tzoulaki⁴, Sudeshna Das², Deborah Blacker², Mark Albers²

¹Kaiser Permanente and UC Berkeley, United States of America

²Mass General Brigham, United States of America

³Haifa University, Israel

⁴Imperial College London, United Kingdom

Alzheimer's disease, the most common type of dementia, affects 6.7 million Americans and costs \$345B annually. Since disease-modifying therapies are limited, repurposing FDA-approved drugs may offer an alternative, expedited path to preventing dementia. Hypertension is a major risk factor for dementia onset. However, prior observational studies contrasting antihypertensive drug classes (Angiotensin Converting Enzyme inhibitors: ACEI, Angiotensin Receptor Blockers: ARB, and Calcium Channel Blockers: CCB), provided mixed results.

We hypothesize that ACEI have an off-target pathogenic mechanism. To test this assumption, we emulate a target trial comparing patients initiating ACEI vs ARB using electronic health records from the US Research Patient Data Registry. We perform intention-to-treat analyses among 25,507 patients aged 50 and over, applying Inverse Propensity score of Treatment Weighting to balance the two treatment arms and accounting for the competing risk of death.

In a cause-specific Cox Proportional Hazards (PH) model, the hazard of dementia onset was higher in ACEI vs ARB initiators (HR=1.10 [95% CI: 1.01-1.21]). Findings were robust to outcome model structures (ie, Cox PH vs nonparametric) and generalized to patients with no hypertension diagnosis at initiation but receiving such drugs for another indication (e.g., heart failure).

Ongoing work includes evaluating differential effects by brain penetrance, discovering subgroups of responders, and assessing the mediating role of blood pressure (BP) control with ACEI vs ARB. Future research will incorporate longitudinal markers (e.g., BP, HbA1c, LDL) in time-to-event models and consider stroke incidence or recurrence under ACEI vs ARB initiation as a mediator.

38: Optimal Utility-Based Design of Phase II/Phase III Programmes with Different Type of Endpoints in the Setting of Multiple Myeloma

Haotian Wang¹, Peter Kimani¹, Michael Grayling², Josephine Khan², Nigel Stallard¹

¹Warwick Clinical Trials Unit, United Kingdom

²Johnson & Johnson Innovative Medicine, United Kingdom

Background High failure rates in phase III oncology trials, often due to overoptimistic assumptions based on limited phase II information, highlight the significant costs and risks associated with drug development. This underscores the importance of approaches that effectively link phase II and phase III trials, balancing resource allocation and decision-making to ensure phase III trials are appropriately powered to optimise success rates.

Method We propose a novel method to determine the optimal phase II sample size that maximizes overall utility of the successful programme. The method evaluates go/no-go decision criteria between phase II and phase III based on phase II outcomes including strategy of choosing the optimal go/no-go threshold, calculating the expected phase III sample size, and ensuring the desired power for the entire programme. Existing methods¹ enable optimal designs when the same time-to-event endpoint is used in both phase II and phase III. But in practice, survival data are often not reliably observed in phase II. Our method allows binary outcome data obtained from phase II to inform the sample size calculation for the phase III trial that will use a correlated time-to-event endpoint.

Results The proposed method is illustrated by application in multiple myeloma, using achieving minimal residual disease as the endpoint in phase II and progression free survival (PFS) as the endpoint in phase III. With initial parameters set according to MAIA trial², we found

the optimal utility and corresponding optimal phase II sample size. We also did sensitivity analysis under different scenarios based on the change of response and treatment related parameters, the value of the go/no-go decision threshold, the prior distribution of response rate and utility-related parameters such as benefits obtained after approval. Our method would provide the optimal design and also give an expected utility of the whole phase II and phase III programme.

Reference 1. Kirchner, M., Kieser, M., Götte, H. & Schüler, A. Utility-based optimization of phase II/III programs. *Stat. Med.* **35**, 305–316 (2016).

2. Facon, T. et al. Daratumumab plus Lenalidomide and Dexamethasone for Untreated Myeloma. *N. Engl. J. Med.* **380**, 2104–2115 (2019).

39: Beyond First Events: Advancing Recurrent Adverse Event Estimates in Clinical Research.

Nicolas Sauvageot, Leen Slaets, Anirban Mitra, Zoe Craig, Jane Gilbert, Lilla Di Scala, Stefan Englert

Johnson & johnson, Switzerland

Safety analyses of adverse events (AEs) are critical for evaluating the benefit-risk profile of therapies; however, these analyses often rely on simplistic estimators that fail to fully capture the complexity present in safety data. The SAVVY consortium, a collaboration between pharmaceutical companies and academic institutions, aims to improve the estimation of the probability of observing the first AE by time t , using survival techniques appropriately dealing with varying follow-up times and competing events (CEs). Through simulation studies¹ and a meta-analysis², the project demonstrated that common methods for estimating the probability of first events such as incidence proportions, Kaplan–Meier (KM) estimators, and incidence densities often fail to account for important factors like censoring and CEs. It concluded that the Aalen–Johansen estimator is the gold standard when focusing on the first event, providing the most reliable estimates, particularly in the presence of CEs.

Only considering first events does not reflect the real burden that a patient may experience in clinical studies. Nevertheless, usual safety reporting and existing research predominantly focuses on the first AE, overlooking the recurrent nature of AEs. Recognizing that both first and subsequent events provide a more accurate representation of safety profiles, there is a clear need to describe both first- and recurrent-AEs in safety reporting.

The objective of this work is to identify appropriate methods for analyzing recurrent AEs in the presence of varying follow-up times and CEs. To achieve this, we perform a simulation study within a recurrent event framework to compare several estimators quantifying the average number of events per subject over time, including:

- Event Rate
- Exposure Adjusted Event Rate (EAER)
- Mean Cumulative Count (MCC) without accounting for CEs
- MCC accounting for CEs³

Our simulations evaluate the performance of these methods regarding bias and examine the impact of various trial characteristics such as the proportion of censoring, the amount of CEs, the AE rate, and the evaluation time point. We illustrate and further strengthen the simulation-based results using real clinical trial data.

- References**
- 1: Stegherr R et al. Estimating and comparing adverse event probabilities in the presence of varying follow-up times and competing events. *Pharm Stat.* 2021 Nov;20(6):1125-1146.
 - 2: Rufibach, K et al. Survival analysis for AdVerse events with VarYing follow-up times (SAVVY): summary of findings and assessment of existing guidelines. *Trials* 25, 353 (2024).
 - 3: Dong H et al. Estimating the burden of recurrent events in the presence of competing risks: the method of mean cumulative count. *Am J Epidemiol.* 2015 Apr 1;181(7):532-40.

Causal Inference for Improved Clinical Collaborations: A Practicum

2025-08-28 09:15 - 13:00, Biozentrum U1.111

Organizers: Alex Ocampo, Cristina Sotto and Jinesh Shah in collaboration with the PSI special interest group in Causal Inference.

Objective

Causal inference is emerging as an indispensable tool for statisticians to properly answer clinical questions of interest. This is due to a mathematically rigorous framework - i.e. potential outcomes - that can explicitly formalize causal effects (estimands) of interest and their identification assumptions. An often-overlooked benefit of adopting a causal toolkit is that it can help create a bridge between statisticians and subject matter experts. For example, causal diagrams can visualize the interplay between various clinical factors and the paths on these diagrams can be used to identify effects of interest together with clinical colleagues. Additionally, causal effects can be defined with simple contrasts of potential outcomes which are generally more closely related to clinical questions than the parameters of statistical models.

This mini symposium will equip participants with fundamental tools from causal inference to enable them to improve their collaborations with clinicians and other non-statistician subject matter experts. Through an introductory lecture on causal inference, a guided hands-on practicum in small breakout groups, and a panel discussion with causal inference experts, attendees of this mini symposium will have the chance to experience how causal inference can assist in improving collaborations between statisticians and clinicians.

Agenda

Session I (09:15 – 10:45)

Chair: Alex Ocampo

09:15 – 09:20 **Welcome** (Alex Ocampo)

09:20 – 09:45 **Introduction to Causal Inference** (Giusi Moffa)

A bite-size introduction will be presented to give participants an overview of the concepts, tools, and language of causal inference.

09:45 – 10:45 **Case Study Practicum** (in breakout groups)

Participants will have the chance to brainstorm and apply causal thinking to a real case study in small groups.

These case studies were drafted by the following statisticians who are applying these tools in clinical trials: Stefan Englert, Lilla Di Scala, Tim Morris, Yannis Jemiai, Cristina Sotto, Alex Ocampo, Jinesh Shah

Coffee Break (10:45 – 11:30)

Session II (11:30 – 13:00)

11:30 – 12:00 **Case Study Overviews**

The context of the case studies will be presented by their respective contributors with some time for comments from the breakout groups.

12:00 – 13:00 **Panel Discussion**

The mini-symposium will conclude with a panel discussion of causal inference experts who will provide their experiences using causal tools in clinical collaborations. Participants will also have a chance to ask questions relevant to the mini-symposium and causal inference more generally.

Panelists: Erica Moodie, Giusi Moffa, Antonio Remiro-Acózar, Stijn Vansteelandt, Emily Granger

Early Career Biostatisticians' Day

2025-08-28 09:15 - 13:00, ETH E21 & E23

Organizers: The ISCB Early Career Biostatisticians' (ECB) Committee and the Local Organising Committee (LOC).

Student Gathering

Please join us for our annual networking event at Markthalle on Sunday, 24th August 2025 at 6pm! This will be a great opportunity to network and connect with other students and early career biostatisticians. This event is free of charge to anyone attending the main conference or ECB Day.

ECB Day

The ECB Day focuses on discussing challenges faced by statisticians and conducting biostatistical research. We welcome all ISCB conference participants to attend regardless of their career stage. In previous years, the topics covered have included working in public health vs industry or academia, working in statistics consultancy, effectively planning and organising a project, navigating professional relationships, work-life balance, and ethical challenges.

Invited Speakers

The following biostatisticians will share their experiences and lessons learned while working as biostatisticians.

Professor Maria Grazia Valsecchi

Senior Professor in Medical Statistics, Bicocca Centre of Bioinformatics, Biostatistics and Bioimaging, University of Milano-Bicocca, Milan, Italy

The key role of biostatisticians in producing methods and applications that improve clinical research and scientific evidence.

Clinical research covers many areas: it evaluates the effectiveness of therapeutic and health-care interventions, the accuracy of diagnostic procedures, the role of new biomarkers, the performance of prognostic or predictive models and many other aspects related to health. Clinical research, if carried out rigorously and efficiently, provides timely results that have

a direct impact on clinical practice, patient care and eventually public health. Research is characterized by multidisciplinarity and biostatistics plays an important role, contributing to all phases of its development: from the definition of the clinical/biomedical question, to the design of the study, the collection and statistical analysis of data, and finally to the proper documentation and communication of the results obtained. For this reason, the profession of biostatistician, or medical statistician, is exciting and interesting, since it implies a role as a scientist, a person who gets to the heart of the research content, contributes with good and innovative methods to produce original data, guarantees the methodological rigor that is necessary for deriving scientific evidence. The biostatistician ethical code of behaviour is also fundamental to preserve the integrity of research for the benefit of subjects involved in the study and of those who will be treated in the future according to the findings. In the presentation I will show, through my work experience as a medical statistician, how exciting it is the interplay between applied and methodological research and how important it is the contribution of our discipline in the production of better research and scientific evidence.

Dr. Karen Lamb

Associate Professor, University of Melbourne

Effective communication strategies for biostatisticians to establish and sustain successful collaboration

Effective communication with collaborators is critical for a successful career in biostatistics. Although an essential skill, communication is rarely emphasised in university training for statisticians. Where offered, communication courses typically focus on written or oral presentation skills. While useful, these courses overlook effective strategies for day-to-day communication required by statisticians. In this presentation, I will discuss essential communication skills biostatisticians need and provide tips and examples of how I use these skills in my own work. I will describe how to establish collaborator trust, how to guide communication pathways, and how to value the experience you bring to a collaboration.

Invited Speakers

Solomon Beer

PhD Student, University of Galway

In recent years there have been a number of opportunities for PhD students as part of a cohort in a targeted subject area, such as Science Foundation Ireland's Centres for Research Training (CRT) and the UK Research and Innovation's Centres for Doctoral Training (CDT). These centres generally include an initial training period where PhD students with a diverse range of relevant academic backgrounds are introduced to theory, methods and application in the centre's subject area. There is a wide range in the research focus for these centres,

from machine learning and data science to renewable energy and environmental science. I am a third year PhD student in the fourth and final cohort of a CRT in Genomics Data Science, which has students spread across six universities in Ireland, and I will discuss my experience of the PhD journey as part of this programme, including some of the positives and negatives that myself and my peers have found through studying for a PhD as part of a cohort.

Workshop: Dilemma Game

In another change from previous years, this year's ECB Day will also include an interactive, workshop-style session in which all participants will be able to get involved in discussions about the challenges faced by statisticians and conducting biostatistical research. For this, we will be using The Dilemma Game app, developed by Erasmus University Rotterdam (EUR), where we pose questions and scenarios relating to professionalism and integrity in research for discussion. You can learn more about The Dilemma Game here: <https://www.eur.nl/en/about-university/policy-and-regulations/integrity/research-integrity/dilemma-game> We recommend that all ECB day attendees download the app prior to attending the mini-symposium.

iOS: <https://apps.apple.com/us/app/dilemma-game/id1494087665>

Android: <https://play.google.com/store/apps/details?id=nl.eur.dilemmagame&hl=nl>

Speed Networking

This session will see attendees divided into groups of about ten. Each participant will have 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.

Schedule

Session I (09:15 – 10:45)

Chair: Autumn O'Donnell

9:15 - 9:20: Opening address – ECB Chair – Autumn O'Donnell

9:20 - 9:55: Invited Speaker: Dr. Karen Lamb

9:55 - 10:10: Student Speaker: Solomon Beer

10:10 - 10:45: Speed Networking

Coffee Break (10:45 - 11:30)

Session 2 (11:30 - 13:00)

Chair: Davide Paolo Bernasconi

11:30 - 12:20: Workshop: Dilemma Game

12:20 - 12:55: Prof. Maria Garzia Valsecchi

12:55 - 13:00: Closing address

Statistical Research needs to improve – on the important roles of simulation studies and guidance for analysis

2025-08-28 09:15 - 13:00, Biozentrum U1.131

Organizers: Anne-Laure Boulesteix (Munich, Germany), Willi Sauerbrei (Freiburg, Germany)

Objective

Although new biostatistical methods are published at a very high rate, many of these developments are not independently evaluated, raising potential concerns about the accuracy and validity of the results. Similar to the well-known phases of research in drug development, Heinze et al. (2024) propose to identify four phases of methodological research. The first of the four phases (I) covers proposing a new methodological idea while providing, for example, logical reasoning or proofs. The three further phases aim at providing empirical evidence helping to evaluate the method's performance in either simulations studies or real-world analyses, with a gradually increasing complexity of the settings and level of evidence. Phase II considers a narrow target setting, while phase III relies on an extended range of settings and for various outcomes, accompanied by appropriate application examples. Finally, phase IV involves investigations that establish a method as sufficiently well-understood to know when it is preferred over others and when it is not; including a systematic exploration of its potential pitfalls.

In the first session, we will have four talks starting with an introductory presentation of the phases concept, followed by three presentations, each revisiting the development history of a specific important biostatistical method, in light of this concept. These three talks will aim at illustrating what phases of methods' development and evaluation were considered and how they were implemented. Together, they may contribute to a further refinement of the phases of methodological research and stimulate discussions around these pivotal issues. In the second session we will have four talks from TG3 (Carsten Schmidt), TG5 (Rima Izem), the open science panel (Sabine Hoffmann) and about a joint project of TGs 2 and 4 (Aris Perperoglou).

Heinze, G., Boulesteix, A. L., Kammer, M., Morris, T. P., White, I. R., & Simulation Panel of the STRATOS Initiative. (2024). Phases of methodological research in biostatistics—building the evidence base for new methods. Biometrical Journal, 66(1), 2200222.

Program

Session I (09:00 – 10:30) Phases of methodological research

Chair: Anne-Laure Boulesteix

09:00 – 09:20 How Biostatistical Methods Mature: Understanding the Four Phases of Methodological Research

Georg Heinze (Medical University of Vienna, Austria)

09:20 – 09:40 Phases of development of the Weighted Cumulative Exposure modeling

Michał Abrahamowicz (McGill University, Montreal, Canada)

09:40 – 10:00 Phases of development of the Multivariable Fractional Polynomial Interaction (MFPI) approach
newline Willi Sauerbrei (Medical Center - University of Freiburg, Germany)

10:00 – 10:20 Will Net Benefit trump Net Reclassification Index as a measure for incremental value of markers in prediction models? A historical perspective

Ewout Steyerberg (University Medical Center Utrecht, Netherlands)

10:20 – 10:30 Discussion

Session II (11:00 – 12:30) Talks from TGs and panels

Chair: Willi Sauerbrei

11:00 – 11:23 Mission impossible? Specifying target estimands for long-term risks and benefits of novel therapies

Rima Izem (Novartis, Basel, Switzerland) for TG5

11:23 – 11:45 An overview on recent works and activities of the STRATOS topic group TG3 “Initial data analysis”

Carsten Oliver Schmidt (University of Greifswald, Germany) for TG3

11:45 – 12:08 An overview and recent developments of the STRATOS Open Science panel

Sabine Hoffmann (Ludwig-Maximilians-Universität Munich, Germany) for the open science panel

12:08 – 12:30 Adjusting for Covariate Measurement Error in Non-Linear Regression: Comprehensive Phase 2 Results from the STRATOS TG2-TG4 Study

Aris Perporoglou (GSK, London, UK) for TG2/TG4

Program Abstracts

Session I

How Biostatistical Methods Mature: Understanding the Four Phases of Methodological Research

Georg Heinze¹, Anne-Laure Boulesteix², Michael Kammer³, Tim P. Morris⁴, Ian R. White⁴

¹Medical University of Vienna, Center for Medical Data Science, Institute of Clinical Biometrics, Vienna, Austria

²LMU Munich, Institute for Medical Information Processing, Biometry and Epidemiology, Munich, Germany

³Medical University of Vienna, Department of Medicine III, Division of Nephrology, Vienna, Austria

⁴UCL, MRC Clinical Trials Unit, Institute of Clinical Trials and Methodology, London, UK

Similar to the well-known phases of research in drug development, Heinze et al. [1] identified four phases of methodological research in biostatistics. The initial phase (I) centers on the development of a novel method. It typically raises the rationale for the method's relevance and novelty, and includes theoretical justifications or formal mathematical proofs. Basic illustrations or toy examples may be contained, but comprehensive empirical evaluation is usually not yet considered. This phase is often limited to a single publication presenting the new idea.

In Phase II, an initial evaluation in a controlled and limited simulation setting is performed. Typically, the method's properties are tested under ideal or simplified conditions, and using well-defined, specific data structures. This phase may include an illustrative real data example and provide a first software implementation, but generalizability is not yet the focus. Many journal articles with biostatistical contributions could be assigned to this phase, but few of them make it to the next phase.

Phase III comprises broad evaluations of a method across diverse settings to investigate a method's wider applicability. This usually includes simulation studies or example applications that span over various settings (e.g., different sample sizes, effect sizes, distributions, sometimes even different types of outcome variables). Alternative methods are well-selected based on evidence, and comparisons among methods are often conducted as neutral comparison studies, avoiding or at least disclosing possible biases. This phase helps in identifying strengths and weaknesses of methods across multiple use cases.

The final phase IV provides meta-methodological insights of a method already in use by practitioners (other than its inventors): it further clarifies when and why the method works well or poorly. It may comprise a narrative or systematic review of simulation results, or wide

comparative performance studies, sometimes largely extending the originally intended fields of application. After that phase, a method is recognized as mature, enabling recommendations for or against its use in specific contexts or according to potential users' level of statistical knowledge and experience. Finally, guidance documents or tutorials for applied users may emerge.

In this introductory talk I will discuss these and some further aspects of the classification and the impact it may have on different stakeholders, such as methodologists, applied researchers, reviewers and journal editors, and funders and policy makers.

1. Heinze, G., Boulesteix, A. L., Kammer, M., Morris, T. P., White, I. R., & Simulation Panel of the STRATOS initiative (2024). Phases of methodological research in biostatistics-Building the evidence base for new methods. *Biometrical Journal*, **66**(1), e2200222. <https://doi.org/10.1002/bimj.202200222>

Phases of development of the Weighted Cumulative Exposure modeling

Michał Abrahamowicz¹

¹McGill University, Montreal, Canada

Weighted Cumulative Exposure (WCE) methodology has been developed to allow for flexible modelling of the cumulative effects of time-varying exposures (TVE) [1]. In time-to-event analyses, the joint impact of past TVE values for person i is quantified as

$$\text{WCE}_i(u) = \sum_t w(u - t) [X_i(t)],$$

where u is the current time when the hazard is assessed, and $X_i(t)$, $t < u$, represent TVE values observed at earlier times. The essential component of the model is the weight function $w(u - t)$, which is estimated using cubic splines and indicates how the importance of the TVE value observed at time t , for the hazard at time u (where $u > t$), varies with time since it was measured [1].

The WCE modeling, originally developed for Cox proportional hazards analyses [1], has been extended to competing risks, marginal structural models, and mixed effects linear modeling of longitudinal changes in a quantitative outcome [2].

The talk will present an overview of the phases of the development and establishing of the WCE methodology, including its consecutive extensions, and validation in simulations. I will discuss how and to what extent our work on the WCE modelling followed the phases identified by Heinze et al. in their recent paper on the phases of methodological research in biostatistics [3]. In addition, further phases such as (a) establishing the need for the new

methodological development, (b) proof-of-concept phase, and (c) refining the estimation and statistical inference, will be outlined.

In this context, three important aspects of real-world applications will be briefly discussed. (i) Firstly, I will emphasize the need to incorporate substantive knowledge, and the related challenges. (ii) Secondly, I will illustrate the ability of the WCE analyses to provide new insights into, and generate new hypotheses about, the underlying biological processes linking the exposure with the outcomes. (iii) I will also outline how some real-world results stimulated new methodological developments, necessary to address additional analytical challenges.

Finally, the need of future research to carry out the additional phase, focusing on neutral simulations to further validate the WCE methodology and compare it with the existing alternative approaches, as recommended in [3], will be briefly presented.

1. Sylvestre M-P & Abrahamowicz M. Flexible modeling of the cumulative effects of time dependent exposures on the hazard. *Statistics in Medicine*. 2009 Nov;28(27):3437-3453.
2. Abrahamowicz M. Assessing cumulative effects of medication use: new insights and new challenges. Invited Commentary. *Pharmacoepidemiology and Drug Safety*. 2024 Jan;33(1):e5746. doi: 10.1002/pds.5746.
3. Heinze G., Boulesteix AL, Kammer M., Morris TP, White I. *Phases of methodological research in biostatistics*. Biometrical Journal. 2024.

Phases of development of the Multivariable Fractional Polynomial Interaction (MFPI) procedure

Willi Sauerbrei¹, Patrick Royston²

¹Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center - University of Freiburg, Freiburg, Germany

²MRC Clinical Trials Unit at UCL, Institute of Clinical Trials and Methodology, University College London, London, UK

MFPI is an extension of the well-established Multivariable Fractional Polynomial (MFP) approach to regression modelling. MFPI was formulated in the context of RCTs to look for an interaction with a continuous variable. The linear interaction model is the simplest special case. More generally, the aim is to investigate for an interaction of a categorical variable with a continuous variable in the framework of a regression model [1]. In the clinical context (RCTs), the key component of this method is the continuous treatment effect function (TEF). We used the bootstrap to perform stability analyses of such functions [2]. Our procedure was inspired by the STEPP (Subpopulation Treatment Effect Pattern Plot) approach. The latter was in vogue some 25 years ago to investigate possible interactions in breast cancer research [3]. We compared the approaches in some examples [4].

Regarding selection of the specific functions, we initially suggested four approaches with varying flexibility (FLEX1 to FLEX4). The details are demonstrated in a Stata paper in which we also compared MFPI with STEPP [5]. Using a large simulation study, we showed the advantages of MFPI over categorization-based methods. Regression splines were also considered as competitors and did not yield much better results. No other spline approaches were investigated [6,7].

We proposed a strategy to average several functions [8] which allowed us to conduct meta-analyses for a continuous variable. Using IPD data from eight RCTs in breast cancer, we illustrated several methodological issues relating to averaging the eight TEF functions [9].

1. Royston, P. and Sauerbrei, W. (2004): A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. *Statistics in Medicine*, 23:2509-2525.
2. Sauerbrei, W. and Royston, P. (2007). Modelling to extract more information from clinical trials data: On some roles for the bootstrap. *Statistics in Medicine*, 26(27), 4989-5001.
3. Bonetti, M. and Gelber, R.D. (2000): A graphical method to assess treatment-covariate interactions using the Cox model on subsets of the data, *Statistics in Medicine* 19: 2595–2609.
4. Sauerbrei, W., Royston, P. and Zapien, K. (2007): Detecting an interaction between treatment and a continuous covariate: a comparison of two approaches. *Computational Statistics and Data Analysis*, 51: 4054-4063.
5. Royston, P. and Sauerbrei, W. (2009): Two techniques for investigating interactions between treatment and continuous covariates in clinical trials. *The Stata Journal*, 9: 230-251.
6. Royston, P. and Sauerbrei, W. (2013): Interaction of treatment with a continuous variable: simulation study of significance level for several methods of analysis. *Statistics in Medicine*, 32(22):3788-3803.
7. Royston, P. and Sauerbrei, W. (2014): Interaction of treatment with a continuous variable: simulation study of power for several methods of analysis. *Statistics in Medicine*, 33: 4695-4708.
8. Sauerbrei, W. and Royston, P. (2011): A new strategy for meta-analysis of continuous covariates in observational studies. *Statistics in Medicine*, 30(28):3341-3360.
9. Sauerbrei, W., & Royston, P. (2022). Investigating treatment-effect modification by a continuous covariate in IPD meta-analysis: an approach using fractional polynomials. *BMC medical research methodology*, 22(1), 98.

Will Net Benefit trump Net Reclassification Index as a measure for incremental value of markers in prediction models? A historical perspective

Ewout Steyerberg¹, Ben Van Calster² for STRATOS TG6

¹University Medical Center Utrecht, Netherlands

²KU Leuven, Belgium

Intro: Markers such as lab measurements and omics features hold promise to improve predictions for individual patients. Various measures can be used to quantify the incremental value of such markers. We aim to place two relatively recent measures in historical perspective: Net Benefit (NB) and Net Reclassification Index (NRI).

Methods: The NB was introduced by Vickers & Elkin in 2006 [1], and Net Reclassification Index (NRI) by Pencina et al in 2008 [2]. Both papers have high citations rates (total >4000 and >6000; in 2024: 553 and 295 respectively). Both measures can consider the situation that a reference prediction model is extended with a covariate, either categorical or continuous ('marker extension').

Results: The NB fits in the line of research on utility measures, where true positive (TP) classifications usually are weighted as more important than false positive (FP) classifications. NB is weighted sum of TP and FP, with the weight related to the decision threshold to classify patients as high vs low risk. A related measure is Relative Utility, as proposed by Baker [3].

The NRI is a reclassification measure, where higher risk is an improvement for those with an event, and lower risk for those without an event. For binary classification, the sum of NRI for events and NRI for non-events is equal to the improvement in Youden index (difference in sensitivity plus difference in specificity). Remarkably, Youden index and NB were already described in 1884 in a 1 page paper [4]. The NRI has been criticized for various reasons, including statistically improper behaviour (testing and estimation problems) and fundamental limitations (not accounting for consequences of classifications, which may be context-dependent, and a risk of misinterpretation) [5]. The NRI is equal to NB if the decision threshold is the event rate, so can be considered a simplified case of NB.

Conclusion: NRI and NB arose from different research traditions, which were already defined in 1884. NRI did not go through systematic phases of evaluation and should not be a prime performance measure for the performance of markers to classify patients as low versus high risk. Time will tell whether NB will trump NRI.

1. AJ Vickers, EB Elkin. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making* 2006; 26 (6), 565-74
2. MJ Pencina, RB D'Agostino Sr, RB D'Agostino Jr, RS Vasan. Evaluating the added

- predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med* 2008; 27(2), 157-72
3. SG Baker. Putting risk prediction in perspective: relative utility curves. *JNCI* 2009;101:1538–42
 4. CS Peirce. The numerical measure of the success of predictions. *Science*, 1884
 5. M Leening, M Vedder, J Witteman, M Pencina, M Pencina, E Steyerberg. Net reclassification improvement: Computation, interpretation, and controversies: A literature review and clinician's guide. *Ann Intern Med* 2014;160:122-31

Session II

Mission impossible? Specifying target estimands for long-term risks and benefits of novel therapies

Rima Izem¹, Paola Rebora², Nicholas Bakewell³, Mitchell Gail⁴, Suzanne Cadarette⁵ for TG5

¹Statistical Methodology, Novartis Pharma AG, Basel, Switzerland

²School and Medicine and Surgery, University of Milano-Bicocca, Italy

³Health Services Research, University of Toronto, Canada

⁴Biostatistics Branch, National Institutes of Health, Rockville, Maryland, USA

⁵Leslie Dan Faculty of Pharmacy, University of Toronto, Canada

The STRATOS Study Design Topic Group (TG5) aims to offer guidance on planning and designing observational studies. Proper planning, informed by subject-matter expertise, ensures that research objectives are clearly defined, clinically relevant, and that the chosen study design is appropriate and valid. Despite its apparent simplicity, flaws in study design are frequently reported, highlighting the need for robust guidance from this subteam.

One TG5 topic of interest includes the review of main challenges in planning clinical trials or observational studies to answer causal questions about the long-term risks and benefits of treatments for chronic conditions. In chronic care, extended exposure to treatments raises questions about long-term safety or effectiveness, necessitating further studies.

The current practice often involves designing studies to compare the initiation of a new treatment with standard care on long-term outcomes. However, the treatment landscape is dynamic. Patients may experience multiple intercurrent events after initiating a treatment, such as dose escalation, switching treatments, therapy gaps, or concurrent treatments, which can influence outcomes. Ignoring these intercurrent events or censoring follow-up at these events allows estimation but muddies the causal inference. Therefore, estimands often focus on quantifying the effect of treatment initiation at the expense of complex exposure patterns.

A potential alternative that TG5 is exploring is to ask cumulative exposure questions at fixed follow-up periods informed by drug utilization patterns in real-world settings.

**An overview on recent works and activities of the STRATOS topic group TG3
“Initial data analysis”**

Carsten O. Schmidt¹, Marianne Huebner², Lara Lusa³

¹Institute for Community Medicine, University Medicine of Greifswald, Greifswald, Germany

²Department of Statistics and Probability, Michigan State University, East Lansing, MI, USA

³Department of Mathematics, Faculty of Mathematics, Natural Sciences and Information Technology, University of Primorska, Koper, Slovenia

The key principle of Initial Data Analysis (IDA) is to provide reliable knowledge about the data underlying the main statistical analyses (MDA). The STRATOS topic group TG3 “Initial data analysis” aims to improve awareness of IDA as an important part of the research process and to provide guidance on conducting IDA in a systematic and reproducible manner in pursue of transparent and reproducible science. IDA focuses on the workflow from metadata setup, data cleaning, data screening, data quality assessments, reporting prior to conducting the MDA. This talk will provide an overview on these steps and introduces an international effort to develop a statistical analysis plan template in cooperation with all STRATOS topic groups for observational studies that incorporates a systematic IDA plan.

An overview and recent developments of the STRATOS Open Science panel

Sabine Hoffmann¹

¹Institute for Medical Information Processing, Biometry, and Epidemiology, Ludwig-Maximilians-Universität Munich, Germany

The scientific community, publishers and funders are increasingly encouraging open science practices with the idea that “scientific knowledge of all kinds, where appropriate, should be openly accessible, transparent, rigorous, reproducible, replicable, accumulative and inclusive” [1]. The STRATOS Open Science panel was funded to promote open science practices by providing guidance on ways to achieve this idea. This talk will give a general overview of the importance of open science practices in the design and analysis of observational studies in biomedical research and then focus on two ongoing projects concerning guidance on data sharing through synthetic data generation and a project that illustrates how to deal with analytical choices (“researcher degrees of freedom”) in the analysis of observational studies.

1. Parsons S, Flavio Azevedo F, Elsherif MM, Guay S, Shahim ON, Govaart GH, Norris E,

Aoife O'Mahony A, Parker AJ, Todorovic A, et al. A community-sourced glossary of open scholarship terms. *Nature Human Behaviour*, 6(3):312–318, 2022.

Adjusting for Covariate Measurement Error in Non-Linear Regression: Comprehensive Phase 2 Results from the STRATOS TG2-TG4 Study

Aris Perperoglou¹, Mohammed Sedki², Anne Thiébaut³, Michal Abrahamowicz⁴, Paul Gustafson⁵, Victor Kipnis⁶, Laurence Freedman⁷ on behalf of the STRATOS TG2 & TG4 collaborative groups

¹GSK, London, UK

²Université Paris-Saclay, France

³INSERM National Institute of Health and Medical Research, Villejuif, France

⁴McGill University, Montreal, Canada

⁵Department of Statistics, The University of British Columbia, Vancouver, British Columbia, Canada

⁶Biometry Research Group, Division of Cancer Prevention, National Cancer Institute, Bethesda, Maryland, USA

⁷Biostatistics Unit, Gertner Institute for Epidemiology and Health Policy Research, Sheba Medical Center, Tel Hashomer, Israel

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 in such non-linear settings.

This blinded, 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, Imputation/Regression Calibration (MI/RC), and Simulation Extrapolation (SIMEX), combined with flexible modelling techniques (B-splines (BS), P-splines (PS), Fractional Polynomials (FP2), and Natural Splines (NS)). Datasets featured a binary outcome, a continuous covariate with classical error (X^*), and a replicate substudy. The true non-linear functional form, covariate distribution, error variance, and error distribution were initially withheld. Phase 1 used 5 pilot datasets; Phase 2 expanded to 155 unique datasets by varying sample sizes, measurement error (ME) variance, error distribution (Normal, shifted-Gamma), and true functional forms. Performance was assessed by log Mean Absolute Error (logMAE).

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 es-

timated most accurately; J-shaped forms were most challenging. A shifted-Gamma ME distribution yielded slightly better accuracy for most methods. Notably, SIMEX was less sensitive to increased ME magnitude and, unlike MI and Bayes, showed no substantial accuracy improvement with larger replication substudy size.

Enhancing Cancer Clinical Trials with Patient-Reported Outcomes: Insights from SISAQOL-IMI

2025-08-28 13:00 - 17:30, Biozentrum U1.111

Organizers: Corneel Coens on behalf of SISAQOL-IMI

Contributors: Flora Mazerolle (Modus outcome), Saskia Le Cessie (Leiden University Medical Center), Johannes Giesinger (Medical University of Innsbruck), Jammbe Musoro (EORTC HQ)

Objective

The SISAQOL-IMI (Setting International Standards in Analysing Patient-Reported Outcomes and Quality of Life Endpoints in Cancer Clinical Trials) is an international multi-disciplinary consortium with funding from the Innovative Medicines Initiative (IMI) aimed to reach standardization for the design, analysis, interpretation and presentation of patient-reported-outcomes (PROs) in cancer clinical trials across various stakeholders (e.g., industry, regulators, HTA bodies, academia, clinicians, patients).

This mini-symposium will report on the experience and progress that led to 146 specific consensus-based recommendations. We will also demonstrate how these recommendations can be easily implemented following practical validation. Application of these recommendation will ensure that PRO objectives can be translated into meaningful design, analysis and reporting considerations. This symposium is relevant not only for statisticians but also for other stakeholders (such as clinicians, patients, pharmaceutical industry, ...). As a large part of the recommendation are aimed at improving communication and understanding of PRO objectives so that the results of cancer clinical trials can be interpreted by all in a standardized manner.

Presenters

Flora Mazerolle

Flora completed her MSc in Socioeconomic Statistics and Data Processing at University Lumière Lyon 2 (France) and joined Modus Outcomes in 2018. She specializes in the analysis of patient-reported outcomes (PRO) data from clinical trials, including both statistical and psychometric analyses in various therapeutic fields, such as oncology, including various type of cancers and hematologic malignancies. As a statistician, she leads quantitative-oriented

projects and contributes to every aspect of these projects, from analysis specification to programming and reporting.

Saskia Le Cessie

Professor Saskia le Cessie (PhD, Dept. of Clinical Epidemiology and dept. of Biomedical data Sciences) is a broadly oriented statistician, with specific expertise in statistical methods for clinical epidemiological research. Her research in medical statistics and epidemiological methods is generally inspired by collaboration with medical researchers, and she has performed research in different areas including prediction models, competing risks and multistate models, measurement variability, combining control groups in case-control studies, and goodness of fit. The focus of her current research is on epidemiological and statistical methods for non-randomised studies. She is an Associate Editor of Clinical Trials and a member of the STRATOS initiative, a large collaboration of experts in many different areas of biostatistical research aiming to provide accessible and accurate guidance in the design and analysis of observational studies. She is one of the WP 3 leaders on single-arm-trials within the SISAQOL consortium.

Johannes Giesinger

Johannes Giesinger (PhD, MSc), is a research psychologist with main interest in methodological work on PRO measures in various medical fields and holds a degree in biostatistics and epidemiology. From 2013 to 2015 he has been a post-doc fellow at the Netherlands Cancer Institute in Amsterdam and he is currently working as a senior researcher at the Medical University of Innsbruck. As a member of the EORTC QLG since 2009, he is serving in both the Statistical Support Group and Grant Review Committee as well as contributing his scientific expertise to the development of numerous PRO measures. He has also led several international research projects on PRO methodology, is one of the WP6 leaders on clinically meaningful change within the SISAQOL consortium.

Jammbe Musoro

Jammbe joined the EORTC in 2015 after completing a PhD in Biostatistics at the Academic Medical Center, University of Amsterdam. He supports both the Quality of Life and Statistics Departments. Currently, he serves as the statistician for the EORTC Cutaneous Lymphoma and Thyroid Cancer Groups, providing expertise in the statistical design of cancer clinical trials. In his role with the Quality of Life Department, he contributes to the design, analysis, and reporting of studies focusing on quality of life endpoints. Jammbe is actively engaged in quality of life research and leads various projects, including the EORTC Minimally Important Difference project, which aims to establish interpretation guidelines for the EORTC QLQ-C30. He regularly contributes to EORTC educational courses and participates in international collaborative projects like SISAQOL-IMI, which aims to standardize the use, analysis, and interpretation of patient-reported outcome data in cancer clinical trials.

Outline

Patient-reported outcomes (PROs) in cancer clinical trials are used to directly measure the patient experience and are becoming increasingly vital in evaluating treatment risks, benefits, and tolerability. However, the lack of consensus among stakeholders often hampers the interpretation and comparability of PRO data.

SISAQOL-IMI (Setting International Standards in Analysing PROs and Quality of Life End-points in Cancer Clinical Trials) is an international, multidisciplinary consortium funded by the Innovative Medicines Initiative. Launched in 2021, this four-year collaborative project will publish its final recommendations in 2025, establishing standards for the design, analysis, interpretation, and presentation of PROs in cancer clinical trials. Recently, SISAQOL-IMI was honoured by the American Statistical Association (ASA) for its exemplary partnerships among academia, industry, patient representatives, and government organizations.

In this half-day mini-symposium, representatives of SISAQOL-IMI will present practice-changing highlights of the 146 total recommendations. More specifically, the symposium will focus on four key scientific areas:

- Randomised Controlled Trials (RCTs): Discussing how PROs can evaluate the comparative clinical benefit of interventions and describe the patient perspective more broadly.
- Single Arm Trials (SATs): SATs often include PROs to explore patient perspectives, support future PRO-related hypotheses in RCTs, and complement clinician-reported adverse events. SATs may be the best option when RCTs are not feasible.
- Presenting and Visualising PRO Results: Addressing the optimal presentation of PRO data for different stakeholder groups using graphic displays.
- Clinically Relevant Thresholds for PRO Scores: Examining the interpretation of clinically relevant differences and changes in PRO scores, a challenging yet vital aspect due to varying definitions, terminology, and methodologies.

The symposium will conclude with a moderated panel discussion on the practical implementation of the recommendations and how these may impact the use of PRO data in drug development and clinical decision making.

Session 1: Introduction (20 min – Johannes Giesinger)

The absence of clear PRO objectives has led to uncertainty in cancer clinical trials, causing confusion in their analysis and reporting. Current practices often show inconsistent termin-

nology, an inflation of analyses, and potentially conflicting results, hindering decision-making and cross-validation of PRO results across studies. Using estimands that reflect relevant and concise PRO objectives can improve the design, analysis, interpretation, and communication of PRO results. The terminology and framework of the initial SISAQOL recommendations will be presented, including a taxonomy of PRO objectives. We will also demonstrate how the estimand framework was used to bridge the gap from research question to actual analysis.

Session 2: Randomised Clinical Trials (25 min – Jammbe Musoro)

Unobserved data is a persistent issue in PRO analyses due to voluntary patient participation, characterized by intercurrent events and missing data. Intercurrent events affect endpoint interpretation and must be aligned with the intended PRO objective, especially in oncology where events like death are not uncommon. Strategies for handling missing data should consider potential informative relationships and assess robustness. Reporting data availability at each assessment timepoint using standardized completion rates and available data rates is recommended to address selection bias in an uniform way across trials.

Session 3: Single Arm Trials (25 min – Saskia Le Cessie)

Single Arm Trials (SATs) face additional challenges due to the lack of a randomized concurrent control. Specifying PRO objectives using the ICHE9 (R1) estimand framework is crucial as current practices are hampered by unclear objectives. Strategies for handling terminal events like death vary, with the while-alive strategy often preferred for quality of life. Causal inference methods offer new opportunities for generating evidence, but untestable causal assumptions must be plausible and supported by sensitivity analyses. This session will explore these strategies and their application in SATs.

Session 4: Results Communication and Visualization (25 min – Flora Mazerolle)

This session will highlight key SISAQOL-IMI recommendations for communicating PRO results from oncology trials, focusing on visualization. Key principles for transparency include aligning figures with pre-specified objectives. Best practices for including sample size details, missing data, and intercurrent events will be presented, along with a template for integrating intercurrent event data into PRO figures. Strategies for reporting statistical significance and distinguishing confirmatory analyses from exploratory findings will be discussed, aiming to improve the accessibility of PRO results for both expert and non-specialist readers.

Session 5: PRO Score Interpretation Thresholds (25 min – Johannes Giesinger)

The heterogeneous terminology around thresholds for interpreting PRO data poses challenges in selecting such thresholds. SISAQOL-IMI has established a harmonized terminology differentiating types of PRO score interpretation thresholds, focusing on patient- and group-level data. These types of threshold correspond to specific statistical analysis methods to support their correct implementation and interpretation. This session will provide recommendations on reporting the use of PRO score interpretation thresholds in cancer clinical trial data analysis and interpretation.

Panel discussion (30 min, all)

This interactive panel discussion will bring together the speakers to delve into the practical implementation of SISAQOL-IMI recommendations. The panel will include all session presenters and the discussion will focus on the challenges in harmonizing PRO objectives. Speakers will be asked to identify areas where these recommendations are expected to have practice-changing impact. Attendees will have the opportunity to engage with the panelists through a Q&A format, fostering a dynamic exchange of ideas and experiences aimed at improving the utility of PRO data in cancer clinical trials.

