CEN2023 Conference Book

From Data to Knowledge.

Advancing Life Sciences.

5th Conference of the Central European Network (CEN)

V 1.1
August 31, 2023

THANK YOU

to our sponsors and funding partners

Platinum Partners

NOVARTIS

Cytel

Swiss National Science Foundation

Roche

Basel basel.ch

AMGEN

Springer

Gold Partners

Karger

DATAMAP

Boehringer Ingelheim

posit™

BeiGene

DENALI THERAPEUTICS

Bristol Myers Squibb™

sanofi

janssen Pharmaceutical Companies of Johnson & Johnson

Strategic Focus Area Personalized Health and Related Technologies

Silver Partners

CRC Press Taylor & Francis Group

CHAPMAN & HALL

Table of Contents

About the Conference

The 5th Conference of the Central European Network will take place from 3-7 September, 2023, at the Biozentrum of the University of Basel, Switzerland.

The conference theme “From Data to Knowledge. Advancing Life Sciences” highlights the special and important role played by biometricians (statisticians and data scientists) when extracting knowledge from data with the ultimate goal of advancing life sciences. The goal of CEN2023 is to present and discuss recent developments in biometry with its applications in life sciences, medicine, pharmacology, research and development in pharmaceutical industry, environmental statistics, and genomics (see the full list of topics).

CEN2023 is a joint conference of

- the Austro-Swiss Region (ROeS),
- the German Region (DR), and
- the Polish Region

of the International Biometric Society (IBS) and supported by the Basel Biometric Society (BBS).

This year, we offer the possibility for ‘tandem talks’ regarding biostatistics for joint presentations at CEN2023 and the GMDS 2023 conference in Heilbronn.

Committees

Executive Committee

Arne Bathke	University of Salzburg
Werner Brannath	University of Bremen
Frank Bretz	Novartis
Uli Burger	Roche
Malgorzata Graczyk	Poznan University of Life Sciences
Annette Kopp-Schneider	German Cancer Research Center

Scientific Committee

Arne Bathke	University of Salzburg
Andrea Berghold	Medical University of Graz
Jan Beyersmann	University of Ulm
Harald Binder	University of Freiburg
Anne-Laure Boulesteix	University of Munich
Werner Brannath	University of Bremen
Frank Bretz	Novartis
Uli Burger	Roche
Tomasz Burzykowski	Hasselt University
Vanessa Didelez	Leibniz Institute for Prevention Research and Epidemiology, Bremen
Janusz Golaszewski	University of Warmia and Mazury in Olsztyn
Malgorzata Graczyk	Poznan University of Life Sciences
Dominik Heinzmann	Novo Nordisk
Benjamin Hofner	Paul-Ehrlich-Institut
Shu-Fang Hsu Schmitz	Janssen
Annette Kopp-Schneider	German Cancer Research Center
Agnieszka Kubik-Komar	University of Life Sciences in Lublin
Dominic Magirr	Novartis
Valentin Rousson	University of Lausanne
Susanne Strohmaier	Medical University of Vienna

Poster Award Committee

Shu-Fang Hsu Schmitz	Janssen
Kristina Weber	Roche
Marvin Wright	Leibniz Institute for Prevention Research and Epidemiology, Bremen

Local Committee

Frank Bretz	Novartis
Uli Burger	Roche
Lilla Di Scala	Johnson & Johnson
Laurence Giullier	Roche
Tracy Glass	Swiss Tropical and Public Health Institute
Achim Güttner	Novartis
Eliane Imfeld	Novartis
Giusi Moffa	University of Basel
Fred Sorenson	Xcenda
Kristina Weber	Roche
Lukas Widmer	Novartis
Marcel Wolbers	Roche

Call for Papers

Biometrical Journal will publish a Special Issue on the theme "From Data to Knowledge. Advancing Life Sciences". The special issue aims at publishing high quality papers on cutting edge statistical methods, especially for but not limited to research presented at CEN 2023.

The review process will follow the general reviewing principles of *Biometrical Journal* and will be handled by the Special Issue guest editors. Considering the aims and scope of *Biometrical Journal*, the methodology described in the manuscripts should be motivated by interesting and relevant problems, ideally from the life sciences. Purely theoretical works would be less suitable for Biometrical Journal.

All papers have to be submitted via the online submission system of *Biometrical Journal*. In submitting the paper, choose “Yes” to “Is this submission for a special issue?”, and select “CEN 2023” in the Special Issue information drop down menu. For more details click here.

Deadline for submission is December 15, 2023.

General Information

Conference Venue

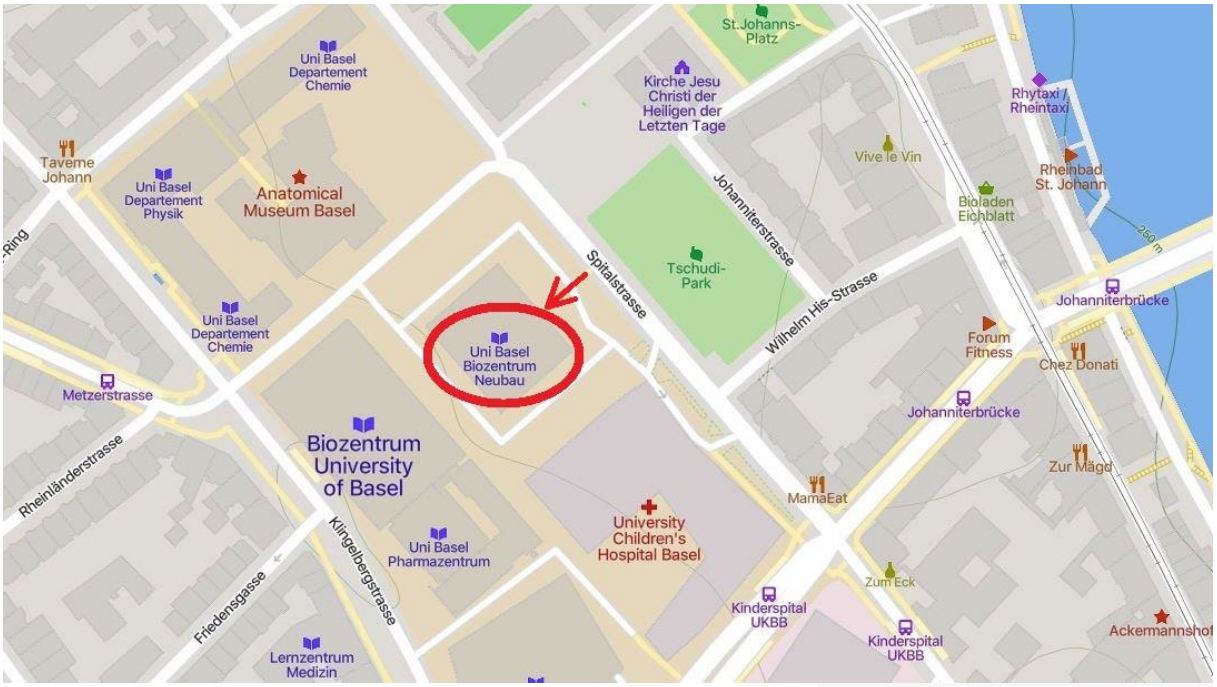
The CEN 2023 conference will take place at the Biozentrum (Neubau) of the University of Basel. The Biozentrum of the University of Basel is located only a few 100 meters from the city center and is easily accessible by public transport, car, bicycle and on foot.

For public transport information, see sbb.ch, the closest stops to Biozentrum are:

- Basel, Kinderspital
- Basel, Johanniterbrücke
- Basel, Metzerstrasse

Address

Biozentrum (Neubau), University of Basel
Spitalstrasse 41
CH - 4056 Basel
Switzerland



© OpenStreetMap, Mapbox and Mapcarta

https://mapcarta.com/W261729542
(Alternative google maps link)

Conference App

The Conference4me smartphone app provides a comfortable tool for planning your participation in the CEN 2023. Browse the complete program directly from your phone or tablet and create your very own agenda on the fly. The app is available for Android and iOS devices.

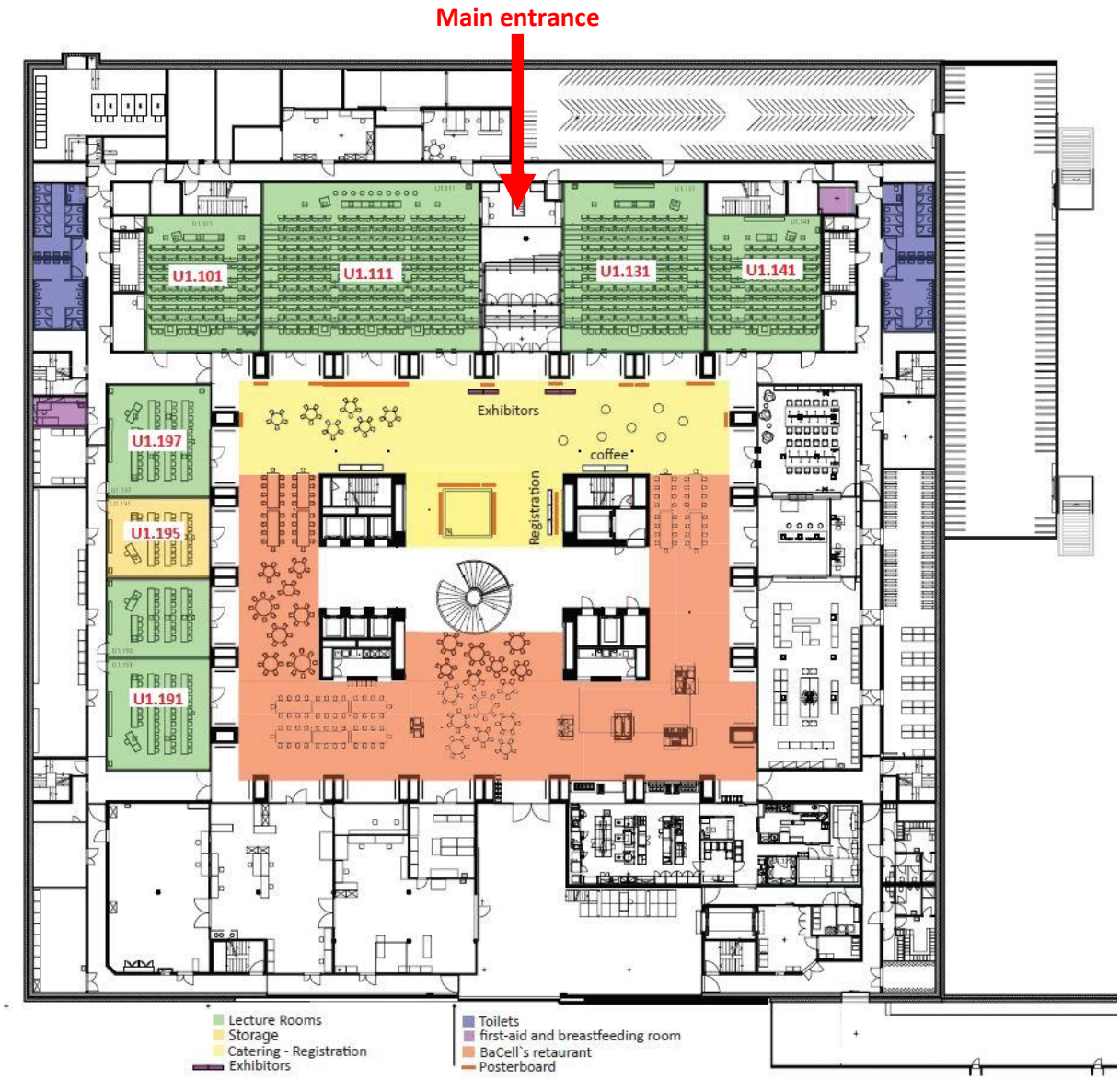
Download the mobile app here, or type “conference4me” in Google Play or iTunes App Store, and select the “CEN 2023” conference in Basel, Switzerland.

Registration

The registration desk at the conference location is open at the following times:

Sunday, September 3	08.00-18.00
Monday, September 4	07.00-18.00
Tuesday, September 5	08.00-18.00
Wednesday, September 6	08.00-14.30
Thursday, September 7	08.00-12.30

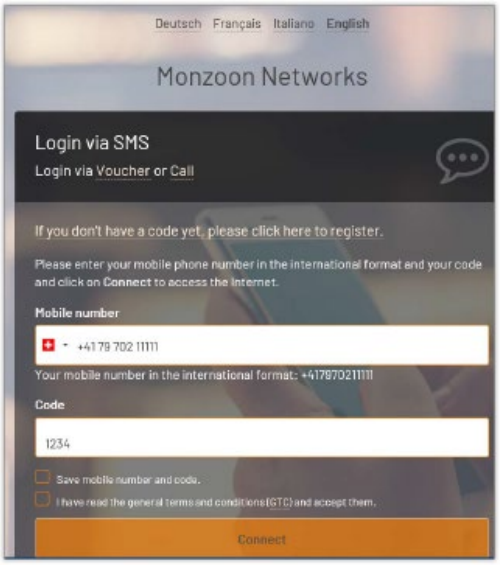
Floor Plan



Internet Access

At the conference venue, you have free access to the University of Basel free Public Wireless LAN (WiFi) by Monzoon Networks.

1. Connect to the to the WiFi “unibas-visitor”.
2. Start your browser and you will be routed automatically to the welcome page.
   (If your browser tries to open an encrypted (https://) webpage you might get an error message. Use then an unencrypted URL to reach the landing page, e.g., http://unibas.ch)
3. Click on “Hier geht’s weiter” (German welcome page) or “Continue here” which brings you to the Monzoon Networks page.
4. Change the language to English.
5. If you already have a code, you can login here with your mobile number and code.



6. Otherwise, click on “If you don't have a code yet, please click here to register” and register your mobile telephone number. You will receive an access code via SMS free of charge.
7. Enter your access code on the welcome page.*
8. Press “Connect” and you will be immediately online.

*Should you not receive a SMS or not have a mobile telephone, you may call the following number from a private Swiss land-line or international mobile telephone: +41 (0) 43 500 3456. You receive a “spoken code” read over the phone (costs for local call within Switzerland occur). This code can be entered under “Code”. For further details please contact Monzoon’s Infoline at 0800 666 966 (7x24).

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.
- Contact Ms Sprüngli from familea at +41 61 260 82 82, who may be able to help with finding additional childcare options.

In case you need further assistance, please contact the organizing team.

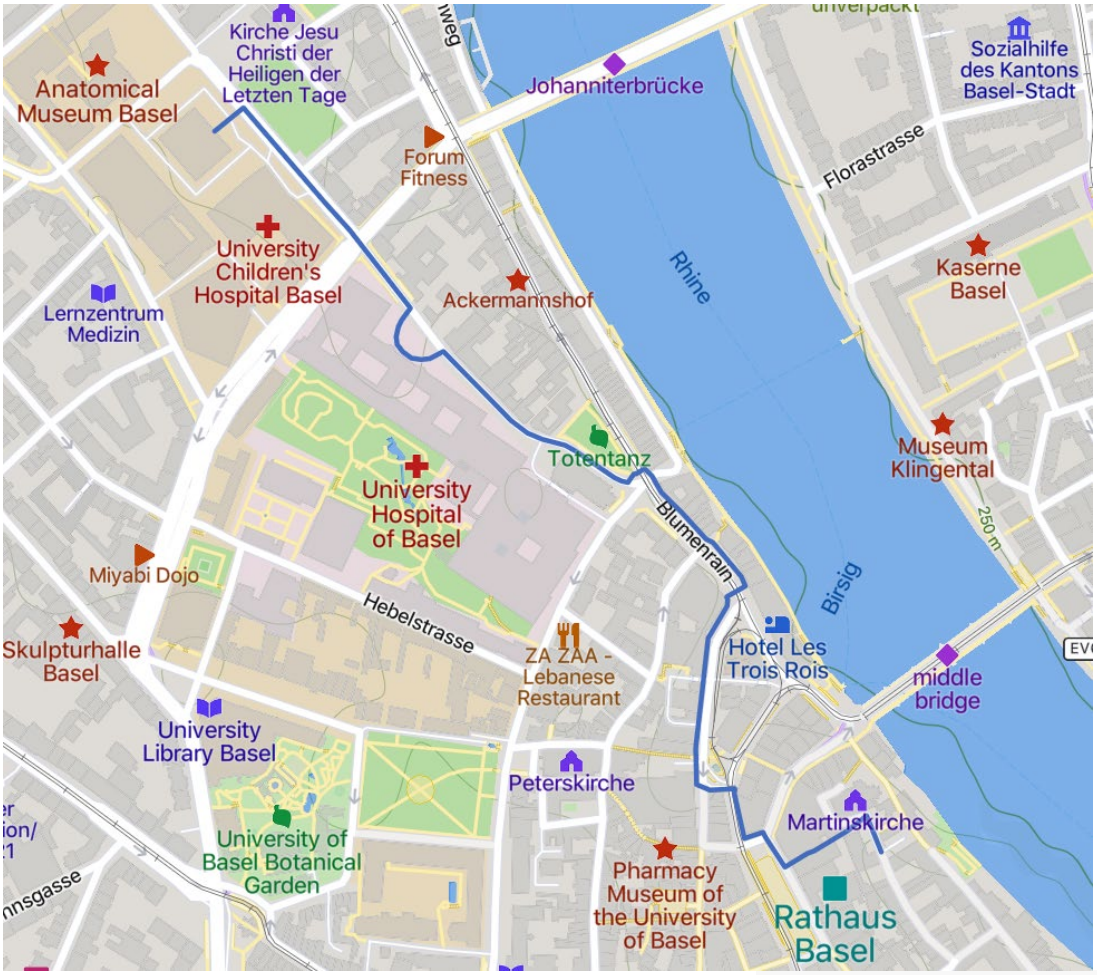
Wine Tasting

An informal Apero with the opportunity to network and taste wine will take place at the conference venue on Monday, September 4, 18:00 – 20:00.

Townhall Reception

The reception will take place on Tuesday, September 5, 18:00 – 20:00, at the city archives behind the parliament building (Rathaus). The address is Martinsgasse 2.

It is approximately a 20-minute walk from the conference venue, see the map for details:



Conference Excursion

The excursion will take place on Wednesday, September 6, 15:00 – 18:00.

Departure

We will depart in buses from the conference venue to the Vitra Campus in Weil am Rhein (Germany) at 15:00. If you want to go to the Vitra Campus by yourself, you can find a travel description at this link. In this case, please make sure that you arrive at Vitra before 15:30.

Walk

We will walk the Rehberger Weg. The Rehberger-Weg, which is around five kilometers long, links two countries, two municipalities, two cultural institutions – and countless stories. The path runs between Weil am Rhein and Riehen, between the Vitra Campus and the Fondation Beyeler. The walk will end at the restaurant of the conference dinner.

Equipment

During the walks the national border is crossed. Please do not forget your **identity card, a bottle of water,** and **sturdy shoes**. We recommend that you download the **Multimedia Guide "24 Stops"** on your mobile which is available on Google Play (for Android devices) and on the Apple App store.

Conference Dinner

The conference dinner will take place on Wednesday, September 6, 18:30 – 22:00 at the Restaurant Landgasthof Riehen, Baselstrasse 38, 4125 Riehen (google maps link).

If you do not join the conference excursion, you can reach the restaurant via Tram 6 (direction "Riehen Grenze") from several central locations in Basel (e.g. Barfüsserplatz, Marktplatz, Claraplatz, or Messe; departures every 8 minutes). The ride takes about 15–25 minutes. Depart from the tram at "Riehen Dorf" and the restaurant will be right next to the tram stop.

To return to your hotel after the conference dinner, you can again use Tram 6 (from tram stop "Riehen Dorf", direction "Allschwil Dorf"). The tram stop is close to the restaurant and trams run every 8 minutes until 22:59 and thereafter every 15 minutes. It stops at several central locations in Basel.

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. The BaselCard app works mostly offline and offers the same free services and discounts. In its electronic version, the guest card also features an interactive city map. For more information, see www.basel.com/en/baselcard.

About Basel

Nestled in the heart of Europe, on both banks of the Rhine, Basel is a pocket-sized metropolis that combines proverbial Swiss quality with a multicultural population. For an introduction to Basel including top places of interest, please have a look at our conference website: cen2023.ch. The website also includes a list of restaurants near the conference location that are open on Sunday (link).

Conference Schedule at a Glance

Sunday, September 3, 2023							
	Lecture Room U1.111	**Lecture Room U1.131**	**Lecture Room U1.141**	**Lecture Room U1.101**	**Seminar Room U1.191**	**Seminar Room U1.197**	**Seminar Room U1.195**
09:00 – 10:30	Short Course 1: Advanced group-sequential and adaptive confirmatory clinical trial designs, with R practicals using rpact	Short Course 2: Bayesian methods for missing covariates in longitudinal studies	Short Course 3: Implementing the estimand framework in global drug development: Application of causal inference approaches	Short Course 4: Go fastR: High Performance Computing with R	IBS Council meeting		
10:30 – 11:00	*Coffee Break*						
11:00 – 12:30	Short Course 1 (part 2)	Short Course 2 (part 2)	Short Course 3 (part 2)	Short Course 4 (part 2)	IBS Council meeting		
12:30 – 14:00	*Lunch*						
14:00 – 15:30	Short Course 1 (part 3)	Short Course 2 (part 3)	Short Course 5: Target Trial Emulation for Causal Inference from Real-World Data	Short Course 7: Model and Algorithm Evaluation in Supervised Machine Learning	IBS Council meeting	Short Course 6: Improving Precision and Power in Randomized Trials by Leveraging Baseline Variables	IBS DR Council meeting
15:30 – 16:00	*Coffee Break*						
16:00 – 17:30	Short Course 1 (part 4)	Short Course 2 (part 4)	Short Course 5 (part 2)	Short Course 7 (part 2)	IBS Council meeting	Short Course 6 (part 2)	IBS DR Council meeting

Color legend:	Short courses	Plenary	Featured	Invited	Topic contributed	Regular

Color legend:	Short courses	Plenary	Featured	Invited	Topic contributed	Regular

Monday, September 4, 2023

	Lecture Room U1.111	Lecture Room U1.131	Lecture Room U1.141	Lecture Room U1.101	Seminar Room U1.191	Seminar Room U1.197	Seminar Room U1.195
08:30 – 09:00	Opening Remarks (Rooms U1.111, U1.131 & U1.141)						
09:00 – 10:00	Keynote Presentation 1 (Rooms U1.111, U1.131 & U1.141) **Ruth Keogh** *Causal inference with observational data: A survival guide*						
10:00 – 10:30	Poster Speed Session 1 (Rooms U1.111, U1.131 & U1.141)						
10:30 – 11:00	*Coffee Break*						
11:00 – 12:40	Session 1 Neutral comparison studies in methodological research	Session 2 Anticipated non-proportional hazards in confirmatory RCTs	Session 3 Prediction models	Session 4 Biometrical Journal Showcase	Session 5 Finding the right dose – Project Optimus and beyond	Session 6 Statistical Modeling I	Session 7 Statistical hypothesis testing
12:40 – 14:00	*Lunch*				Business Meeting: AG Bayes-Methodik	Business Meeting: AG Non-Clinical Statistics	Business Meeting: AG Landwirtschaftliches Versuchswesen
14:00 – 15:40	Session 8 Statistics in Practice 1	Session 9 A causal inference perspective on estimands in clinical trials	Session 10 Personalized health care	Session 11 Safety and benefit/risk assessment in master protocols	Session 12 Critical cross-disciplinary collaboration in dose optimization in oncology	Session 13 Statistical Modeling II	Session 14 Clinical trials I
15:40 – 16:10	*Coffee Break*						
16:10 – 17:50	Session 15 Statistics in Practice 2	Session 16 Causal discovery with a view to the life sciences	Session 17 Prognostic and predictive biomarker in personalized medicine	Session 18 Clinical trials II	Session 19 Oncology trials	Session 20 Statistical Modeling III	Session 21 General topics
18:00 – 20:00	*Wine Tasting*						

Color legend:	Short courses	Plenary	Featured	Invited	Topic contributed	Regular

Tuesday, September 5, 2023

	Lecture Room U1.111	Lecture Room U1.131	Lecture Room U1.141	Lecture Room U1.101	Seminar Room U1.191	Seminar Room U1.197	Seminar Room U1.195
09:00 – 10:00	Keynote Presentation 2 (Rooms U1.111, U1.131 & U1.141) **Alicja Szabelska-Beręsewicz** *Statistical methods to analyse the structure of the microbiome based on cereal leaf beetle (Oulema melanopus) data*						
10:00 – 10:30	Poster Speed Session 2						
10:30 – 11:00	*Coffee Break*						
11:00 – 12:40	Session 22 Net benefit, win odds, and win ratio: Methods, analysis, and interpretation	Session 23 Time-to-event analysis I	Session 24 Machine learning	Session 25 Young Statisticians 1	Session 26 Multiple testing	Session 27 Statistical issues in health care provider comparisons	Session 28 Meta-analysis and systematic reviews I
12:40 – 14:00	*Lunch*		Business Meeting: EFSPI methodology leaders group	Business Meeting: AG Nachwuchs	Business Meeting: AG Nichtparametrische Methoden	Roundtable Discussion: Statistical issues in health care provider comparisons	Business Meeting: AG Population Genetics and Genome Analysis
14:00 – 15:40	Session 29 Generalized pairwise comparisons	Session 30 Time-to-event analysis II	Session 31 Interpretable machine learning in biostatistics: Methods, applications and perspectives	Session 32 Young Statisticians 2	Session 33 Endpoints in clinical trials and medical product development	Session 34 Meta-analysis and systematic reviews II	Session 35 Epidemiology
15:40 – 16:10	*Coffee Break*						
16:10 – 17:50	Session 36 Adaptive designs	Session 37 Causal estimands for time to event data	Session 38 From multivariate to high-dimensional and functional data	Session 39 Young Statisticians Panel - "Should I stay or should I go"	Session 40 Beyond the two-trials paradigm for generating pivotal evidence in drug development	Session 41 Use of external data	Session 42 Simulation studies
18:00 – 20:00	*Townhall Reception*						

Color legend:	Short courses	Plenary	Featured	Invited	Topic contributed	Regular

Wednesday, September 6, 2023

	Lecture Room U1.111	Lecture Room U1.131	Lecture Room U1.141	Lecture Room U1.101	Seminar Room U1.191	Seminar Room U1.197	Seminar Room U1.195
08:30 – 10:10	Session 43 Causal inference	Session 44 Online hypothesis testing and subgroup analyses in complex innovative designs	Session 45 High-dimensional analysis	Session 46 Data Monitoring Committees 1	Session 47 Statistical software engineering in the pharmaceutical industry	Session 48 Sample size considerations	Session 49 Safety evaluations
10:10 – 10:40	*Coffee Break*						
10:40 – 12:20	Session 50 Covariate adjustment in RCTs	Session 51 IBS-DR / IBS-ROeS Award Session	Session 52 Random forests	Session 53 Real-world evidence	Session 54 Software engineering	Session 55 Non-clinical and toxicology studies	Session 56 Data Monitoring Committees 2
12:20 – 13:00	*Lunch*						
13:00 – 15:00	DR Assembly (Room U1.111) and ROeS Assembly (Room U1.131)						
15:00 – 18:00	*Excursion*						
18:30 – 22:00	*Conference Dinner* **Stephen Senn**						

Color legend:	Short courses	Plenary	Featured	Invited	Topic contributed	Regular

Thursday, September 7, 2023

	Lecture Room U1.111	Lecture Room U1.131	Lecture Room U1.141	Lecture Room U1.101	Seminar Room U1.191	Seminar Room U1.197	Seminar Room U1.195
8:30 – 10:10	Session 57 Causal inference and the art of asking meaningful questions	Session 58 Innovative clinical trial designs	Session 59 Volume-outcome relationships in health care	Session 60 Variable selection	Session 61 Statistical strategies in toxicology	Session 62 Epidemic short-term forecasting in real time	Session 63 Analysis of omics data I
10:10 – 10:40	*Coffee Break*						
10:40 – 12:20	Session 64 Estimands	Session 65 Advancing clinical trial design in rare diseases	Session 66 Advanced survival analysis	Session 67 Dieter Hauschke Memorial	Session 68 Design of preclinical experiments	Session 70 COVID-19	Session 69 Analysis of omics data II
12:20 – 13:00	*Lunch*						
13:00 – 14:00	Keynote Presentation 3 (Rooms U1.111 & U1.131) **Peter Bühlmann** *Learning from other Intensive Care Units: can we improve statistical predictions?*						
14:00 – 14:30	Closing Remarks (Rooms U1.111 & U1.131)						
14:30	*Conference End*						
14:30 – 18:00	STRATOS Satellite Symposium (Room U1.111)						

Conference Presentations at a Glance

Monday, September 4, 2023							
	Lecture Room U1.111	**Lecture Room U1.131**	**Lecture Room U1.141**	**Lecture Room U1.101**	**Seminar Room U1.191**	**Seminar Room U1.197**	**Seminar Room U1.195**
08:30 – 09:00	**Opening Remarks (Rooms U1.111, U1.131 & U1.141)**						
09:00 – 10:00	**Keynote Presentation 1 (Rooms U1.111, U1.131 & U1.141) - Ruth Keogh: *Causal inference with observational data: A survival guide***						
10:00 – 10:30	**Poster Speed Session 1 (Rooms U1.111, U1.131 & U1.141)**						
10:30 – 11:00	***Coffee Break***						
11:00 – 12:40	**S1: Neutral comparison studies in methodological research**	**S2: Anticipated non-proportional hazards in confirmatory RCTs**	**S3: Prediction models**	**S4: Biometrical Journal Showcase**	**S5: Finding the right dose – Project Optimus and beyond**	**S6: Statistical Modeling I**	**S7: Statistical hypothesis testing**
11.00	Pitfalls and Potentials in Simulation Studies (Samuel Pawel)	Methods for non proportional hazards in clinical trials: A systematic review (Cynthia Huber)	Calibrating machine learning approaches for probability estimation: a comparison (Max Louis Jansen)	On the logic of collapsibility for causal effect measures (Vanessa Didelez)	Back-fill cohorts in oncology phase I trials (Lukas Schröter)	Modeling the Ratio of Gamma Distributed Random Variables using Frank's Copula (Moritz Berger)	Nonparametric methods for clustered data in the several sample case (Erin Sprünken)
11.20	Against the "one method fits all data sets" philosophy for comparison studies in methodological research (Carolin Strobl)	A simulation-based comparison of statistical methods for time-to-event data analysis under non-proportional hazards (Florian Klinglmueller)	Advanced statistical modelling for polygenic risk scores by incorporating alternative loss functions (Hannah Klinkhammer)	Randomized p-values in replicability analysis (Thorsten Dickhaus)	Towards efficient dose-escalation guidance of multi-cycle cancer therapies (Sebastian Dragos Weber)	Simplifying complex models: deselection for boosting distributional copula regression (Annika Strömer)	Almost Omnibus Nonparametric Inference for Two Independent Samples (Jonas Beck)
11.40	Phases of methodological research in biostatistics - Building the evidence base for new methods (Georg Heinze)		Comparison of classic polygenic scores with machine learning algorithms to predict blood pressure (Tanja K. Rausch)	Missing data imputation in clinical trials using recurrent neural network facilitated by clustering and oversampling (Halimu Haliduola)	Dose Finding Studies for Therapies with Late-Onset Safety and Activity Outcomes (Thomas Jaki)	Random graphical model of microbiome interactions in related environments (Veronica Vinciotti)	Robust ANCOVA for Small Sample Studies (Konstantin Emil Thiel)
12.00	Discussion with panelists Ruth Keogh and Anne-Laure Boulesteix	Panel discussion with Tobias Mütze, Dominic Magirr and Marcia Rueckbeil	Investigating different numbers of variants in polygenic scores using the ALLIANCE cohort (Lisa-Marie Nuxoll)	Missing data: A statistical framework for practice (James Robert Carpenter)	Discussion with panelists Oliver Boix, Lukas Schröter, Sebastian Dragos Weber and Thomas Jaki	Methods of Model selection for models with common parameters (Onur Gül)	Identifying alert concentrations using a model-based bootstrap approach (Kathrin Möllenhoff)
12.20			Similarity as a basis for data pooling - Improving Local Prediction Models Using External Data (Max Behrens)			Comparing statistical methods for analyzing longitudinally measured ordinal outcomes in rare disease settings. (Martin Geroldinger)	Testing for similarity of multivariate mixed outcomes with application to efficacy-toxicity responses (Niklas Hagemann)

Monday, September 4, 2023							
	Lecture Room U1.111	**Lecture Room U1.131**	**Lecture Room U1.141**	**Lecture Room U1.101**	**Seminar Room U1.191**	**Seminar Room U1.197**	**Seminar Room U1.195**
12:40 – 14:00	***Lunch***				**Business Meeting: AG Bayes-Methodik**	**Business Meeting: AG Non-Clinical Statistics**	**Business Meeting: AG Landwirtschaftliches Versuchswesen**
14:00 – 15:40	**S8: Statistics in Practice 1**	**S9: A causal inference perspective on estimands in clinical trials**	**S10: Personalized health care**	**S11: Safety and benefit/risk assessment in master protocols**	**S12: Critical cross-disciplinary collaboration in dose optimization in oncology**	**S13: Statistical Modeling II**	**S14: Clinical trials I**
14:00	Simulation studies as a tool to assess and compare the properties of statistical methods – an overview (Tim Morris and Brennan C. Kahan)	From causal inference with observational data to estimands in RCTs and back! (Vanessa Didelez)	Swiss Personalized Health Network from clinical routine data to FAIR data for research (Sabine Oesterle)	Analysis of safety data with special attention to platform trials and the estimand framework (Ekkehard Glimm)	Designing Dose-Optimization Studies in Cancer Drug Development: Discussions with Regulators (Olga Marchenko)	A transformation perspective on marginal and conditional models (Torsten Hothorn)	Non-reproducibility between phase II and III: Region selection and go/no-go related bias and methods for its correction (Kristin Schultes)
14:20			Personalized diagnosis in suspected myocardial infarction: the ARTEMIS study (Eleonora Di Carluccio)	How to master the challenges of safety and benefit/risk assessment planning? (Jürgen Kübler)	A Roadmap for Novel Oncology Dose Finding Designs (Revathi Ananthakrishnan)	Mixed-effects Additive Transformation Models with the R Package tramME (Balint Tamasi)	Beyond the two-trials rule (Leonhard Held)
14:40		The danger of extrapolation in RCTs and how to avoid it (Hege Michiels)	Developing a predictive model with causal considerations for the risk for antibiotics resistance based on patient health records (Anat Reiner-Benaim)	Safety and Benefit-Risk Evaluation: Master Protocols for Efficient Evidence Generation (Alessandro Vagheggini)	Pragmatic and Holistic Approach for Dose Finding and Optimization in Oncology Drug Development (Jiang Liu)	A tool to detect nonlinearity and interactions in generalized regression models (Nikolai Spuck)	Dual Primary Endpoints – innovative idea or avoidable risk? (Nele Henrike Thomas)
15:00		What is the role of causal thinking in global drug development? (Mouna Akacha)	Predictors for chronic opioid use – real world evidence using insurance claims data from Switzerland (Ulrike Held)	Discussion with panelist William Wang	Panel discussion with Daniel Li	Uncertainty Estimation in Nonlinear Models within the Profile Likelihood Framework (Tim Litwin)	Novel weighted approach for estimating effects in principal strata with missing data in RCTs (Dominik Heinzmann)
15:20		Discussion by Oliver Dukes	Individual-specific network inference for prediction modelling: a plasmode simulation study (Mariella Gregorich)			Bayesian nonlinear functional subspace shrinkage with application to gene expression dose-response data (Julia Christin Duda)	A RCT assessing the protective efficacy of an odour-based 'push-pull' malaria vector control strategy in reducing human-vector contact (Adrian Eugen Denz)
15:40 – 16:10	***Coffee Break***						

Monday, September 4, 2023							
	Lecture Room U1.111	**Lecture Room U1.131**	**Lecture Room U1.141**	**Lecture Room U1.101**	**Seminar Room U1.191**	**Seminar Room U1.197**	**Seminar Room U1.195**
16:10 – 17:50	**S15: Statistics in Practice 2**	**S16: Causal discovery with a view to the life sciences**	**S17: Prognostic and predictive biomarkers in personalized medicine**	**S18: Clinical trials II**	**S19: Oncology trials**	**S20: Statistical Modeling III**	**S21: General topics**
16:10	Continuation of first session (S8: Statistics in Practice 1)	Causal discovery: Benchmarking algorithms and Bayesian analyses in the life sciences (Jack Kuipers)	Confident and Logical Selection of the Cut-point of a Biomarker for Patient Targeting (Yang Han)	How patient preference studies can support decision-making in early drug development (Phases 1-2) (Sheila Dickinson)	Improved treatment effect estimation for time-to-event outcomes in subgroups displayed in forest plots based on shrinkage methods (Mar Vázquez Rabuñal)	Optimal Subsampling Design for Polynomial Regression (Torsten Reuter)	Biostatistics/Biometrics for physicians – essential or unnecessary? How do practicing physicians and dentists evaluate biostatistics? A cross-sectional survey (Maren Vens)
16:30			Bayesian hierarchical models for biomarker discovery in drug combination screens (Manuela Zucknick)	Assessment of pharmacokinetic linearity after repeated drug administration (Alexander Bauer)	Deconstructing PFS to understand the treatment effect and predict OS in oncology drug development (Francois Mercier)	Optimal design for identifying alert concentrations (Kirsten Schorning)	How to estimate parameters in weighing design? (Małgorzata Graczyk)
16:50		Consistent and efficient mixed integer programming for causal discovery (Ali Shojaie)	Using knockoffs for controlled predictive biomarker identification (Kostas Sechidis)	Assurance of three-way PK (PD) Biosimilarity studies with multiple coprimary endpoints (Rachid El Galta)	Modeling the dose-response relationship using advanced tumor metrics (Cornelia Ursula Kunz)	Bias through endogenous time-varying covariates in the analysis of cohort stepped-wedge trials: a simulation study (Jale Basten)	On estimation of use efficiency (Jens Hartung)
17:10		Causality-inspired ML: what can causality do for ML? (Sara Magliacane)	Logic respecting efficacy measures in the presence of prognostic or predictive biomarker subgroups (Hong Sun)	Real World Evidence Application in Biosimilar Development (Samriddhi Buxy Sinha)	Assessment of the treatment effect in dose ranging studies with time to event endpoints accounting for the intercurrent event of dose reductions (Arunava Chakravartty)	Estimating the conditional distribution in functional regression problems (Thomas Kuenzer)	Performance of point and interval estimators for average sequential attributable fraction - A simulation study (Carolin Malsch)
17:30		Bayesian inference of causal graphs: where we are and where we should go (Mikko Koivisto)	Searching for treatment effect modifiers in manual therapy: Three case studies (Werner Vach)	Quantification Of Dataset Similarity For Small Sample Sizes (Maryam Farhadizadeh)	Practical advice on the reporting of statistical items in the new CONSORT extension for early phase dose-finding trials (Jan Rekowski)	Using Item response theory for testing assumptions underlying clinical scores (Daniel Schulze)	Studying global alien species invasions between 1880 and 2005 with relational event models (Ernst C. Wit)
18:00 – 20:00	***Wine Tasting***						

Tuesday, September 5, 2023							
	Lecture Room U1.111	**Lecture Room U1.131**	**Lecture Room U1.141**	**Lecture Room U1.101**	**Seminar Room U1.191**	**Seminar Room U1.197**	**Seminar Room U1.195**
09:00 – 10:00	**Keynote Presentation 2 (Rooms U1.111, U1.131 & U1.141)** **Alicja Szabelska-Beręsewicz:** ***Statistical methods to analyse the structure of the microbiome based on cereal leaf beetle (Oulema melanopus) data***						
10:00 – 10:30	**Poster Speed Session 2**						
10:30 – 11:00	***Coffee Break***						
11:00 – 12:40	**S22: Net benefit, win odds, and win ratio**	**S23: Time-to-event analysis I**	**S24: Machine learning**	**S25: Young Statisticians 1**	**S26: Multiple testing**	**S27: Statistical issues in health care provider comparisons**	**S28: Meta-analysis and systematic reviews I**
11:00	Net Benefit, Success Odds, and Win Ratio for Non-Censored Observations (Edgar Brunner)	Oncology clinical trial design based on a multistate model that jointly models PFS and OS (Kaspar Rufibach)	Bayesian Uncertainty Quantification in Deep Generative Models for Synthesis of Tabular Medical Data (Patric Tippmann)	Evaluating cancer screening programmes using survival analysis (Bor Vratanar)	multiCASANOVA - Multiple group comparisons for non-proportional hazard settings (Ina Dormuth)	Comparing implants, hospitals and surgeons: lessons learned from the Swiss National Hip & Knee Joint Registry SIRIS (Christian Brand)	LFK index does not reliably detect bias in meta-analysis (Guido Schwarzer)
11:20	Use and interpretation of the net treatment benefit, success odds, and win ratio for censored and non-censored data (Marc Buyse)	Non-Markov non-parametric estimation of complex multistate outcomes after hematopoietic stem cell transplantation (Judith Vilsmeier)	Combining Boosting with Neural Networks for Structuring Latent Representations of Single-Cell RNA-Sequencing Data (Niklas Brunn)	A two-step approach for analysing time-to-event data under non-proportional hazards (Jonas Elias Brugger)	Simultaneous confidence intervals for an extended Koch-Röhmel design in three-arm non-inferiority trials (Martin Scharpenberg)	Comparing procedural quality indicators of health care across regions using restricted mean survival time (Hana Šinkovec)	A REML method for the evidence-splitting model in network meta-analysis (Hans-Peter Piepho)
11:40		Analysis of Time to Treatment Responses: An Application of a Multi-State Model using Semi-Markov Process (Lillian Yau)	One-stage and two-stage detectors comparison in the task of pollen grains recognition (Elżbieta Kubera)	Sample size recalculation for a skewed outcome in two-stage three-arm sequential non-inferiority trials (Maria Vittoria Chiaruttini)	Control of essential type I error rates in clinical trials with multiple hypotheses (Werner Brannath)	Analysing PROM based quality of care indicators in care centers (Els Goetghebeur)	Random-effects meta-analysis of subgroup specific effects and treatment-by-subgroup interactions (Renato Valladares Panaro)
12:00	Inferential methods for generalized pairwise comparisons of censored data (Vaiva Deltuvaite-Thomas)	A general estimation framework for multistate survival processes with flexible specification of the transition intensities (Alessia Eletti)		Modelling antibody kinetics – A systematic review and study design considerations (Stefan Embacher)	Statistical calibration for infinite many future values in linear regression (Lingjiao Wang)	Patient surveys for assessing medical treatment quality (Felix Weidemann)	Nonproportional Hazards in Network Meta-Analysis: Efficient Strategies for Model Building and Analysis (Anna Wiksten)
12:20	Semiparametric and Nonparametric Methods for Covariate Adjustment for GPC Effect Sizes for Multiple Outcomes (Olivier Thas)	Penalized likelihood estimation of stratified semi-parametric Cox models under partly interval censoring (Jun Ma)		Statistical methods for the analysis of massspectrometry data with multiple membership (Mateusz Staniak)	Online multiple testing with heterogeneous data (Sebastian Doehler)	Discussion with panelist Werner Vach	Statistical considerations on the coverage probability of a CI when sequentially combining n-of-1 studies in a cumulative meta-analysis (Eleonora Carrozzo)

Tuesday, September 5, 2023							
	Lecture Room U1.111	**Lecture Room U1.131**	**Lecture Room U1.141**	**Lecture Room U1.101**	**Seminar Room U1.191**	**Seminar Room U1.197**	**Seminar Room U1.195**
	Lunch		**Business Meeting: EFSPI methodology leaders group**	**Business Meeting: AG Nachwuchs**	**Business Meeting: AG Nichtparametrische Methoden**	**Roundtable Discussion: Statistical issues in health care provider comparisons**	**Business Meeting: AG Population Genetics and Genome Analysis**
14:00 – 15:40	**S29: Generalized pairwise comparisons**	**S30: Time-to-event analysis II**	**S31: Interpretable machine learning in biostatistics**	**S32: Young Statisticians 2**	**S33: Endpoints in clinical trials and drug development**	**S34: Meta-analysis and systematic reviews II**	**S35: Epidemiology**
14:00	Multivariate Outcomes and the Need for Generalized Pairwise Comparisons (Arne Bathke)	A parametric additive hazard model for time-to-event analysis (Dina Voeltz)	Interaction difference test for prediction models (Thomas Welchowski)	Enhancing replicability of exploratory variable selections based on clinical trial data (Manuela Rebecca Zimmermann)	Endpoint Development and Analysis Planning in Clinical Trials (Hsien-Ming James Hung)	IPW-based publication bias adjustment in network meta-analysis with clinical trial registries (Ao Huang)	State space models as a flexible framework for monitoring epidemics (Thomas Hotz)
14:20	Visuals for Generalized Pairwise Comparisons: innovative tools to explore treatment effects on multiple prioritized outcomes (Samuel Salvaggio)	Pseudo-Value Regression Trees (Alina Schenk)	Multi-Objective Counterfactual Explanations (Susanne Dandl)	A flexible framework for interpretable and individualized reporting of model results (Hannah Kümpel)		Implementation of the anchor-based indirect comparison method for equivalence margin derivation in biosimilar development (Claudia Hemmelmann)	Pertussis in Belgium - The challenge of using historical serial serological survey data (Sereina Herzog)
14:40	Applications of generalized pairwise comparisons and rank-based procedures in small samples: Bootstrap and permutation tests (Frank Konietschke)	A non-parametric proportional risk model to assess a treatment effect in an application to long-term carcinogenicity assays (Lucia Ameis)	Interpreting Neural Networks: A Biostatistical Perspective (Niklas Koenen)	Confirmatory studies in methodological statistical research: concept and illustration (F. Julian D. Lange)	Design and Analyis of Desirability of Outcome Ranking in Clinical Trials (Toshimitsu Hamasaki)	Imputation of informatively missing data in meta-analyses (Christine Macare)	Estimating the effects of hypothetical behavioral interventions on overweight/obesity incidence using observational data (Claudia Börnhorst)
15:00	Generalized pairwise comparisons as a pragmatic alternative to non-inferiority trial designs (Mickaël De Backer)	Modelling tree survival for investigating climate change effects (Nicole H Augustin)	Explainability of machine learning models for survival analysis: current state and challenges (Mateusz Krzyziński)	Analysis of the Effect of Hyperparameters on Variable Selection in Random Forests (Lea L. Kronziel)	Beyond Proportional Hazards: Multi-Parameter Approaches and Confirmatory Multiple Testing (Martin Posch)	Prognostic models for disease progression in people with multiple sclerosis (Begum Irmak On)	Adverse health outcomes among people with atopic eczema (Julian Matthewman)
15:20	Individualized Net Benefit estimation and meta analysis using generalized pairwise comparisons in N-of-1 trials (Joris Giai)	Developing a survival prediction model – a case study (Samuel Kilian)	Panel discussion	The best of two worlds? A systematic comparison of time-to-event model implementations between R and Python (Lukas Klein)	Discussion by Robert Hemmings	Investigating the heterogeneity between "study twins" (Christian Röver)	Integrated transcriptome- and proteome-wide association studies nominate causal determinants of kidney function (Pascal Schlosser)
15:40 – 16:10	***Coffee Break***						

Tuesday, September 5, 2023							
	Lecture Room U1.111	**Lecture Room U1.131**	**Lecture Room U1.141**	**Lecture Room U1.101**	**Seminar Room U1.191**	**Seminar Room U1.197**	**Seminar Room U1.195**
16:10 – 17:50	**S36: Adaptive designs**	**S37: Causal estimands for time to event data**	**S38: From multivariate to high-dimensional and functional data**	**S39: Young Statisticians Panel – "Should I stay or should I go"**	**S40: Beyond the two-trials paradigm in drug development**	**S41: Use of external data**	**S42: Simulation studies**
16:10	Practical learnings from futility analyses in Phase 3 trials (Gian-Andrea Thanei)	On the choice of estimands in clinical trials with time-to-event outcomes (Mats Stensrud)	Functional data analysis on the example of analysis of variance (Łukasz Smaga)	Panel discussion on "Going Abroad During Your Academic Career". Panelists: Anne-Laure Boulesteix Annika Hoyer Antonella Mazzei Christian Müller Silke Szymczak	Comparison of one-trial and two-trial paradigms in drug assessment (Stella Jinran Zhan)	Estimation of treatment effects in early phase randomized clinical trials involving external control data (Heiko Götte)	Towards more practically relevant method comparison studies by generating simulations based on a sample of real data sets (Christina Nießl)
16:30	Group sequential methods for the Mann-Whitney parameter (Claus Nowak)	Treatment effect measures in clinical trials with time-to-event outcomes: it is time to apply estimand thinking (Tobias Mütze and Vivian Lanius)			When convention meets practicality: Combined analysis testing under the two-trials convention (Dong Xi)	Robust incorporation of external information in two-arm trial hypothesis testing (Silvia Calderazzo)	A simple-to-use R package for mimicking study data by simulations (Giorgos Koliopanos)
16:50	Adaptive selection of binary composite endpoints and sample size reassessment based on blinded data (Marta Bofill Roig)	Estimands for time to event data: a regulator's view (Andreas Brandt)	MultiFANOVA: Multiple Contrast Tests for Functional Data (Merle Munko)		Combining clinical trials to generate pivotal evidence – case studies and reflections (Marc Vandemeulebroecke)	Augmenting randomized trials with real-world data: a simulation study evaluating methods for hybrid control arm analyses (Rafael Sauter)	Comparison of methods for quantifying similarity of datasets (Marieke Stolte)
17:10	Adaptive enrichment designs for clinical trials with multiple endpoints (Koko Asakura)	Discussion with panelist Kaspar Rufibach	Quantile-based MANOVA: A new tool for inferring multivariate data in factorial designs (Marléne Baumeister)		Discussion with panelists Kit Roes, Nigel Stallard, and Hsien-Ming James Hung	Dynamic borrowing to minimize mean squared error and inference with Bayesian bootstrap (Jixian Wang)	How to Simulate Realistic Survival Data? A Simulation Study to Compare Realistic Simulation Models (Maria Thurow)
17:30	A flexible simulation framework of Bayesian adaptive designs (Dominique-Laurent Couturier)		Testing Hypotheses about Correlation Matrices in General MANOVA Designs (Paavo Sattler)			Partial extrapolation in pediatric drug development using robust meta-analytic predictive priors, tipping point analysis and expert elicitation (Florian Voß)	Simulation study to compare methods to analyze time-to-event endpoints in trials with delayed treatment effects (Rouven Behnisch)
18:00 – 20:00	***Townhall Reception***						

Wednesday, September 6, 2023							
	Lecture Room U1.111	**Lecture Room U1.131**	**Lecture Room U1.141**	**Lecture Room U1.101**	**Seminar Room U1.191**	**Seminar Room U1.197**	**Seminar Room U1.195**
08:30 – 10:10	**S43: Causal inference**	**S44: Online hypothesis testing and subgroup analyses in complex innovative designs**	**S45: High-dimensional analysis**	**S46: Data Monitoring Committees 1**	**S47: Statistical software engineering in the pharmaceutical industry**	**S48: Sample size considerations**	**S49: Safety evaluations**
8:30	Generalizing the intention-to-treat effect of an active control from historical placebo-controlled trials to an active-controlled trial (Oliver Dukes)	Online error rate control for platform trials (David Robertson)	Over-optimism in gene set analysis: How does the choice of methods and parameters influence the detection of differentially enriched gene sets? (Milena Wünsch)	Introduction to DMCs: basic principles and some statistical issues (Tim Friede)	First year of the Software Engineering working group - working together across organizations (Daniel Sabanes Bove)	Sample Size Re-estimation for the Wilcoxon-Mann-Whitney and Brunner-Munzel Test (Stephen Schüürhuis)	Considerations of safety estimands and estimators in pivotal and post-market studies (Rima Izem)
8:50	Outcomes truncated by death: a simulation study on the survivor average causal effect (Stefanie von Felten)		Maximum Test Method for the Wilcoxon-Mann-Whitney Test in High-Dimensional Designs (Lukas Mödl)	Regulatory perspective on iDMCs by an experienced iDMC member (Kit Roes)	Refactoring and extending an existing R package across companies - learnings from the crmPack team (Clara Beck)	Sample size recalculation in three-stage clinical trials (Björn Bokelmann)	Long term safety evaluations in the presence of switching: evaluation of two approaches (Sandra Schmeller)
9:10	Target trial emulation avoids bias due to non-alignment at time-zero in studies on site-specific effectiveness of screening colonoscopy (Malte Braitmaier)	Multiple testing of partial conjunction null hypotheses, with application to replicability analysis of high-dimensional studies (Thorsten Dickhaus)	Deriving interpretable thresholds for Variable Importance in Random Forests by permutation (Hannes Buchner)	Visualisation and reporting of safety issues (Rachel Phillips)	To package or not to package - a pragmatic approach to deciding whether an R package is the right solution for your problem and alternatives to consider (Kevin Kunzmann)	Sample size calculations for cluster randomised trials using assurance (Sarah Faye Williamson)	Adverse event burden score as an alternative approach to quantify and compare adverse event burden in clinical trials (Bartosz Jenner)
9:30	Estimating and interpreting causal effects under violation of positivity (Maria Geers)	Graphical procedures for online error control (Lasse Fischer)	Mind your zeros: accurate p-value approximation in permutation testing with applications in microbiome data analysis (Stefanie Peschel)	Behind the scenes: the Data Analysis Center for Data Monitoring Committees (Benjamin Esterni)	Discussion with panelists Wilmar Igl, Thomas Jaki and Gregory Chen	May the power be with you? Influence of sample size calculation on replication success. (Collazo Anja)	Signal detection of adverse drug reactions: The Bayesian power generalized Weibull shape parameter test (Julia Alexandra Dyck)
9:50		Multi-stage adaptive enrichment designs with BSSR (Marius Placzek)	A novel approach to Function-on-Scalar Regression (FoSR) for the analysis of Periodic Time-Series (Konrad Neumann)	Sponsor perspective on best practices for Data Monitoring Committees (Gregory Golm)		Researcher Degrees of Freedom in Power Analyses and Sample Size Planning (Nicole Ellenbach)	Evaluation of adverse events in early benefit assessment (Part I): Firth correction for Cox models in the case of zero events (Lars Beckmann)
10:10 – 10:40	***Coffee Break***						

Wednesday, September 6, 2023							
	Lecture Room U1.111	**Lecture Room U1.131**	**Lecture Room U1.141**	**Lecture Room U1.101**	**Seminar Room U1.191**	**Seminar Room U1.197**	**Seminar Room U1.195**
10:40 – 12:20	**S50: Covariate adjustment in RCTs**	**S51: IBS-DR / IBS-ROeS Award Session**	**S52: Random forests**	**S53: Real-world evidence**	**S54; Software engineering**	**S55: Non-clinical and toxicology studies**	**S56:Data Monitoring Committees 2**
10:40	Improving Power in Randomized Trials by Leveraging Baseline Variables (Kelly Van Lancker)	Online multiple testing with FWER control (Lasse Fischer)	Challenge in distinguishing important from informative variables in random forest prediction models (Césaire Fouodo)	EUnetHTA 21 methods guidelines for EU HTA: The good, the bad, and the ugly - an industry perspective. (Sandro Gsteiger)	Comparing R libraries with SAS's PROC MIXED for the analysis of longitudinal continuous endpoints using MMRM (Gonzalo Duran-Pacheco)	"Lots of time to think": Statistical Consultancy at Ciba-Geigy in the Early 1980s (Andrew Grieve)	Interactive workshop on communicating data to Data Monitoring Committees (Tobias Mütze and David Lawrence)
11:00		Blinded sample size reestimation in clinical trials with time-to-event outcomes based on flexible parametric models (Tim Mori)	Identifying different tree types based on clustering in random forests (Björn-Hergen Laabs)	Quantifying and comparing impact of different sources of uncertainty in the analysis of electronic health records (Maximilian Mandl)	An overview of R software tools to support simulation studies: towards standardizing coding practices (Michael Kammer)	Literature review of dose-response analyses in toxicology (Franziska Kappenberg)	
11:20	Organizing a Data Challenge on Covariate Adjustment in RCTs (Dominic Magirr)	Confounder adjustment with random forests based on local residuals in genetic association studies (Annika Swenne)	Evaluation of network-guided random forest for disease gene discovery (Jianchang Hu)	On selecting a parametric model to predict long-term survival to support health technology assessment (Gregory Chen)	A Julia Package for Bayesian Optimal Design of Experiments (Ludger Sandig)	Hurdles and Signposts on the Road to Virtual Control Groups in Toxicity Studies (Lea A.I. Vaas)	
11:40	Participating in a data challenge on covariate adjustment in RCTs (Craig Wang)	Cluster-robust estimators for multivariate mixed-effects meta-regression (Thilo Welz)	Random Survival Forests for Competing Events: A Subdistribution-Based Approach (Charlotte Behning)	Rule-based estimation of lines of therapy (LoT) from oncological registry data: the SAKK 80/19 AlpineTIR registry (Alfonso Rojas Mora)	Online sample size calculator (Robin Ristl)	The application of prediction intervals in pre-clinical statistics and toxicology using the R package predint (Max Menssen)	
12:00	Panel discussion.	Optimal treatment regimes assisted by algorithms (Mats Stensrud)	Generative modeling of epidemiological data using adversarial random forests (Jan Kapar)	Extended Excess Hazard Models for Spatially Dependent Survival Data (André Victor Ribeiro Amaral)	A multi-platform PDR technology to efficiently build complex tables & reports, flexibly iterate with stakeholders, and easily maintain across workflows (Ming Zou)	Improving Production Capacity and Asset Utilization of Biologics Drug Product Lines Through Simulation (Christian Schmid)	
12:20 – 13:00	***Lunch***						
13:00 – 15:00	**DR Assembly & ROeS Assembly**						
15:00 – 18:00	***Excursion***						
18:30 – 22:00	***Conference Dinner* (Stephen Senn)**						

Thursday, September 7, 2023							
	Lecture Room U1.111	**Lecture Room U1.131**	**Lecture Room U1.141**	**Lecture Room U1.101**	**Seminar Room U1.191**	**Seminar Room U1.197**	**Seminar Room U1.195**
8:30 – 10:10	**S57: Causal inference and the art of asking meaningful questions**	**S58: Innovative clinical trial designs**	**S59: Volume-outcome relationships in health care**	**S60: Variable selection**	**S61: Statistical strategies in toxicology**	**S62: Epidemic short-term forecasting in real time**	**S63: Analysis of omics data I**
8:30	Causal inference with competing events (Jessica Young)	Statistical and regulatory lessons learned from the pandemic (Benjamin Hofner)	Minimum Volume Thresholds of the Federal Joint Committee in Germany (Horst Schuster)	Do we need different variable selection procedures depending on the goal of the statistical model? (Theresa Ullmann)	The joint analysis of multiple sources of multiplicity in the evaluation of regulatory toxicology bioassays (Ludwig A. Hothorn)	Collaborative forecasting of COVID-19 in Germany and Poland (Melanie Schienle)	DESpace: a novel analysis framework to discover spatially variable genes (Simone Tiberi)
8:50		Computationally Efficient Basket Trial Designs Based on Empirical Bayes Power Prior Methods (Lukas Baumann)	The Assessment of Volume-Outcome Associations at IQWiG (Claudia-Martina Messow)	A neutral comparison of algorithms to minimize L0 penalties for high-dimensional variable selection (Florian Frommlet)		Improving short term forecasts of COVID-19 incidence with subnational epidemic indicators (Stefan Heyder)	Pathway analysis for multinomial phenotypes (Taesung Park)
9:10	Are E-values too optimistic or too pessimistic? Both and neither! (Arvid Sjölander)	A biomarker-guided Bayesian response-adaptive phase II trial for patients with metastatic melanoma: The Personalized Immunotherapy Platform (PIP)-Trial design (Serigne Lo)	Modelling volume-outcome relationships in health care (Maurilio Gutzeit)	Effects of Influential Points and Sample Size on the Selection and Replicability of Multivariable Fractional Polynomial Models (Willi Sauerbrei)	An integrated data-driven approach for drug safety prediction (Fetene Tekle)	Predicting the unpredictable: the MOCOS large scale agent based epidemic model (Tyll Krueger)	Multi-omics data integration: Does more mean better for predictive modeling? A large-scale benchmark study (Roman Hornung)
9:30	Optimal regimes for algorithm-assisted human decision-making (Aaron Leor Sarvet)	Generating the right evidence at the right time: Principles of a new class of flexible augmented clinical trial designs (Cornelia Dunger-Baldauf)	Relationship between hospital volume and medium-term survival in breast cancer surgical and oncologic treatment in Lombardy -Italy (Anita Andreano)	Post-estimation shrinkage in full and selected linear regression models in low-dimensional data revisited (Edwin Kipruto)	Statistics in a validation process in toxicology (Tina Lang)	Strong effect of testing in containing Covid-19 (Jan Mohring)	Boosting interaction tree stumps for modeling gene–gene and gene–environment interactions (Michael Lau)
9:50	On the choice of estimands when the role of an intermediate variable is of interest (Rhian Mair Daniel)		Discussion with panelist Tim Mathes	High-Dimensional Variable Selection for Competing Risks with Cooperative Penalized Regression (Lukas Burk)	The Comet assay in vivo – a review of known properties and new findings (Timur Tug)	Multi-step immunity mechanism in ICM UW epidemic agent-based model (PDYN 1.5) (Jędrzej Nowosielski)	Testing for associations in genomic data with distances and kernels: From unconditional to conditional settings (Fernando Castro-Prado)
10:10 – 10:40	***Coffee Break***						

Thursday, September 7, 2023							
	Lecture Room U1.111	**Lecture Room U1.131**	**Lecture Room U1.141**	**Lecture Room U1.101**	**Seminar Room U1.191**	**Seminar Room U1.197**	**Seminar Room U1.195**
10:40 – 12:20	**S64: Estimands**	**S65: Advancing clinical trial design in rare diseases**	**S66: Advanced survival analysis**	**S67: Dieter Hauschke Memorial**	**S68: Design of preclinical experiments**	**S70: COVID-19**	**S69: Analysis of omics data II**
10:40	Asking the right questions when assessing overall survival in a randomized clinical trial that allows for cross over (Silvia Colicino)	Innovations in Clinical Development in Rare Diseases in Children and Adolescents (Robert Beckman)	Modelling chronic disease mortality by methods from accelerated life testing (Marina Zamsheva)	How to assess bioequivalence of two drugs? (Iris Pigeot)	Designs for the simultaneous inference of concentration-response curves (Leonie Schürmeyer)	Why are different estimates of the effective reproductive number so different? A case study on COVID-19 in Germany (Johannes Bracher)	High-dimensional graphical models varying with multiple external covariates (Louis Dijkstra)
11:00	Estimands in practice: revisiting endpoint definitions in the light of the lymphoma patient journey. (Emmanuel Zuber)		A sensitivity analysis approach for the causal hazard ratio in randomized and observational studies (Rachel Axelrod)	What can bioequivalence studies teach us about clinical trials? (Stephen Senn)	D-optimality in preclinical dose-response studies (Leonie Hezler)	On the Impacts of the COVID-19 Pandemic on Mortality: Lost Years or Lost Days? (Valentin Rousson)	Detecting interactions in High Dimensional Data using Cross Leverage Scores (Sven Teschke)
11:20	Data-generating models of longitudinal continuous outcomes and intercurrent events to evaluate estimands (Marian Mitroiu)	Hybrid controlled clinical trials using concurrent registries in Amyotrophic Lateral Sclerosis: A feasibility study (Ruben van Eijk)	Consequences of omitted covariates on treatment estimates in propensity score matched studies (Alexandra Strobel)	On intelligent use of the 3-am gold standard design with test treatment, placebo and active control (Joachim Roehmel)	Applying 10 simple rules for good research practice in pre-clinical research (Philip Jarvis)	Collaborative nowcasting of COVID-19 hospitalization incidences in Germany (Daniel Wolffram)	The impact of missing SNPs in the calculation of polygenic scores (Hanna Brudermann)
11:40	Decentralized clinical trials: scientific considerations through the lens of the estimand framework (Nikolaos Sfikas)	Advancing clinical trial design for small populations: balance and/or rigor? (Kit Roes)	Individual heterogeneity in humoral immune response: A Bayesian frailty approach (Steven Abrams)	Identification of minimal effective dose MED (resp. no observed effect concentration NOEC) in unbalanced designs with possible heterogenous variances (Ludwig Hothorn)	Designing preclinical experiments using factorial and Bayesian approaches (Andreas Allgöwer)	Systematic review on prevention and testing strategies for COVID-pandemic control in economic comparison (Noah Alessandro Castioni)	Pre-processing and quality control of whole genome sequencing data: a case study using 9000 samples from the GENESIS-HD study (Andreas Ziegler)
12:00	An estimand framework to guide model and algorithm evaluation in predictive modelling (Max Westphal)	Panel discussion	Discussion with panelist Steven Abrams	When safety data meet survival analysis (Claudia Schmoor)	Potential of Generalized Pairwise Comparisons in pre-clinical studies (Johan Verbeeck)	Bayesian Poisson Regression and Tensor Train Decomposition Model for Learning Mortality Pattern Changes during Pandemic (Wei Zhang)	More than meets the eye: Dimension reduction and temporal patterns in time-series single-cell RNA-sequencing data (Maren Hackenberg)
12:20 – 13:00	***Lunch***						
13:00 – 14:00	**Keynote Presentation 3 (Rooms U1.111 & U1.131)** **Peter Bühlmann:** ***Learning from other Intensive Care Units: can we improve statistical predictions?***						

14:00 – 14:30	**Closing Remarks (Rooms U1.111 & U1.131)**						
14:30	***Conference End***						
Thursday, September 7, 2023							
	Lecture Room U1.111	**Lecture Room U1.131**	**Lecture Room U1.141**	**Lecture Room U1.101**	**Seminar Room U1.191**	**Seminar Room U1.197**	**Seminar Room U1.195**
14:35 – 16:05	**STRATOS Satellite Symposium – Session 1 (Room U1.111)**						
	Experience and progress with developing guidance for the analysis of key topics in observational research (Willi Sauerbrei and James Carpenter) Initial data analysis plans are part of research projects (Marianne Huebner) Level 1 guidance on conducting and reporting sensitivity analyses for missing data (James Robert Carpenter)						
16:20 – 17:20	**STRATOS Satellite Symposium – Session 2 (Room U1.111)**						
	Statistical analysis of high-dimensional biomedical data: A gentle introduction to analytical goals, common approaches and challenges (Jörg Rahnenführer) Ongoing research towards state-of-the-art in variable and functional form selection for statistical models (Georg Heinze) Data-Driven Simulations to Assess the Impact of Study Imperfections In Real-World Time-to-Event Analyses (Michal Abrahamowicz)						
17:30-18:00	**STRATOS Satellite Symposium – Session 3: Methodological research needs to improve – getting involved to increase future contributions of the STRATOS initiative Location: Lecture (Room U1.111)**						
	Panel discussion with Willi Sauerbrei, Michal Abrahamowicz, Marianne Huebner, Ruth Keogh, James Robert Carpenter (for the STRATOS initiative)						

GMDS Tandem Talks

This year, we offer the possibility for 'tandem talks' regarding biostatistics for joint presentations at CEN2023 and the GMDS 2023 conference in Heilbronn. In total five companion presentations will be given at the GMDS 2023 conference.

1) **Wednesday, 06/Sept/2023 9:50am - 10:10am**
ID: 433 / S49: 5
Presentation Submissions - Regular Session (Default)
Topic of Submission: Time-to-Event Analysis
Keywords: Cox proportional hazards model, Firth correction, adverse events, early benefit assessment

Evaluation of adverse events in early benefit assessment (Part I): Firth correction for Cox models in the case of zero events

Lars Beckmann, Guido Skipka, Anke Schulz
IQWiG, Cologne, Germany; lars.beckmann@iqwig.de

2) **ID: 215 / Poster 1: 5**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Biomarkers and diagnostics, Prevention and handling of missing data
Keywords: diagnostic study, accuracy, missing values, inconclusive, reference standard, index test, sensitivity and specificity, Poster ID M5

Missing values and inconclusive results in diagnostic studies – a scoping review of methods

Katharina Stahlmann[1], Johannes B. Reitsma[2], Antonia Zapf[1]
[1]Institute of Medical Biometry and Epidemiology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany; [2]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands; k.stahlmann@uke.de

3) **Monday, 04/Sept/2023 2:40pm - 3:00pm**
ID: 174 / S14: 3
Presentation Submissions - Regular Session (Default)
Topic of Submission: Clinical trial designs (adaptive and platform designs, external controls, …), Statistical hypothesis testing (covariate adjustment, nonparametric methods, multiple comparisons, …)
Keywords: co-primary endpoints, dual primary endpoints, at-least-one concept, trial design, confirmatory clinical trials

Dual Primary Endpoints – innovative idea or avoidable risk?

Nele Henrike Thomas, Armin Koch, Anika Großhennig
Medical School Hannover, Germany; thomas.nele@mh-hannover.de

4) **Monday, 04/Sept/2023 4:10pm - 4:30pm**
ID: 435 / S21: 1
Presentation Submissions - Regular Session (Default)
Topic of Submission: Free Contributions
Keywords: biostatistics, physicians, dentists, attitude, teaching

Biostatistics/Biometrics for physicians – essential or unnecessary? How do practicing physicians and dentists evaluate biostatistics? A cross-sectional survey

Maren Vens[1], Nina Alida Hartmann[2], Inke Regina König[1]
[1]Institut für Medizinische Biometrie und Statistik, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Universität zu Lübeck, Lübeck, Deutschland; [2]Institut für Sozialmedizin und Epidemiologie, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Universität zu Lübeck, Lübeck, Deutschland; m.vens@uni-luebeck.de

5) **ID: 281 / Poster 1: 15**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Machine Learning and Data Science, Meta-Analysis and Systematic Reviews
Keywords: text classification, machine learning, abstract screening, systemtic review, Poster ID M15

Text classification to automate abstract screening using machine learning

Johannes A. Vey[1], Samuel Zimmermann[1], Maximilian Pilz[1,2]
[1]Institute of Medical Biometry, University of Heidelberg, Germany; [2]Department Optimization, Fraunhofer Institute for Industrial Mathematics (ITWM), Germany; vey@imbi.uni-heidelberg.de

Pre-Conference Short Courses

Advanced group-sequential and adaptive confirmatory clinical trial designs, with R practicals using rpact

Marcel Wolbers[1], **Kaspar Rufibach**[1], **Gernot Wassmer**[2], **Marc Vandemeulebroecke**[3]
[1]Roche Pharma, Basel, Switzerland; [2]rpact and University of Cologne, Germany; [3]Novartis, Basel, Switzerland;
marcel.wolbers@roche.com, kaspar.rufibach@roche.com, gernot.wassmer@rpact.com,
marc.vandemeulebroecke@novartis.com
Keywords: interim analyses; sample size re-calculation; multi-arm multi-stage designs

This course is intended for biostatisticians from pharma and academia who are interested in learning more about advanced topics in group-sequential and adaptive clinical trial designs. Topics covered are the efficient use of interim analyses in group-sequential trials, an introduction to adaptive trials and sample size recalculation, the use of closed testing procedures for adaptive trials with multiple objectives, and multi-arm multi-stage designs. Examples from real clinical trials will be used throughout the presentations. We also aim to discuss operational aspects of implementing such designs in practice.

The course will be a mix of presentations and practicals using the R package rpact, a free and fully validated package for the design and analysis of group-sequential and adaptive trials. We will assume basic familiarity with group-sequential designs and R. Participants are asked to bring a laptop with R and rpact installed. It is the ambition of the instructors to make this course very interactive.

Bayesian methods for missing covariates in longitudinal studies

Nicole Erler[1], **Emanuel Lesaffre**[2]
[1]Erasmus University Medical Center, Rotterdam, the Netherlands; [2]KU Leuven, Belgium; n.erler@erasmusmc.nl,
emmanuel.lesaffre@kuleuven.be
Keywords: Bayesian Methods; Missing Data; Longitudinal Data; Imputation

Missing values commonly complicate the analysis of observational data. Multiple imputation (MI) is considered the "gold standard" for handling incomplete covariates. MI, developed at the beginning of the Computer Age, is based on Bayesian ideas. In complex settings, e.g. involving non-linear associations or multi-level data, the assumptions of the commonly used MI algorithms are, however, often violated, leading to possibly biased results. Thanks to the current computational power, a fully Bayesian approach, allowing us to simultaneously estimate parameters of interest and impute missing values, is now feasible. This approach is theoretically valid and superior to MI in complex settings. Highly complex non-standard missing data models can relatively easily be implemented with the help of freely available software such as the R package JointAI. In this course, we briefly review the essentials of multi-level data, Bayesian concepts and (multiple) imputation. The main focus is on the Bayesian approach to missing values in covariates in multi-level and longitudinal studies, which is motivated and illustrated using examples from clinical and epidemiological studies. Practical sessions will be organized to show the capabilities of the R package JointAI, starting with simpler standard settings and extending to highly complex joint models for longitudinal and survival data and imputation in non-standard settings.

Implementing the estimand framework in global drug development: Application of causal inference approaches

Mouna Akacha[1], **Björn Bornkamp**[1], **Alex Ocampo**[1], **Jiawei Wei**[2]
[1]Novartis Pharma AG, Basel, Switzerland; [2]Novartis Institutes for Biomedical Research Co., Shanghai, China;
mouna.akacha@novartis.com, bjoern.bornkamp@novartis.com, alex.ocampo@novartis.com, jiawei.wei@novartis.com
Keywords: causal inference; conditional estimand; marginal estimand; standardization; ICH E9(R1); hypothetical estimand; principal stratum estimand

This half-day short course introduces how causal inference approaches are relevant and used in the implementation of estimands framework in drug development. It includes 4 lectures:

Lecture 1 - Introduction to Estimands and Causal Inference:

1. Overview of the estimand framework and key points in ICH E9(R1)
2. Introduction to causal inference, including potential outcomes, causal effects and common assumptions;

Lecture 2 - Estimation Methods of Causal Effects Targeting at Hypothetical Estimands:

1. Introduction to common estimation methods, e.g., g-computation, IPW (Inverse probability weighting)
2. RCT examples illustrated using R code;

Lecture 3 - Principal Stratum:

1. Introduction to principal stratum estimand
2. Estimation strategies
3. Case studies in RCTs;

Lecture 4 - Conditional and Marginal Treatment Effects:

1. Introduction to conditional and marginal treatment effects
2. Appropriate estimators for conditional and marginal estimands
3. RCT examples illustrated using R code.

Go fastR: High Performance Computing with R

Michael Mayer[1], **Lukas Widmer**[2]
[1]Posit PBC, Boston, USA; [2]Novartis Pharma AG, Basel, Switzerland; michael.mayer@posit.co, lukas_andreas.widmer@novartis.com

This course will help participants to optimize their R code as well as parallelizing and debugging it on their own machines as well as high-performance computing environments. Example use cases include commonly performed activities for trial design, bootstrapping, cross-validation and related workloads. The following topics will be covered:

Part I: Identifying bottlenecks in your R code, debugging, and optimizing

- Debugging R code & checking correctness
- Profiling R code to identify bottlenecks
- Optimizing bottlenecks locally: packages, vectorizing, logical indexing

Part II: R parallelization on high performance computing environments (HPCE)

- Amdahl's law and limits of achievable speed up
- Parallelizing work onto compute clusters via clusterMQ and batchtools
- Consistently loading packages, .libPaths() and options() on R workers
- Uncorrelated random number generation for parallel R code
- Debugging R code in batchtools and clusterMQ jobs

Part III: Case studies and code examples

- Bootstrapping
- Cross-validation
- Trial simulations under replication
- Within-chain parallelization with several chains in Stan
- Bring your own problem: start to speed-up your own code with the help of the instructors.

Target Trial Emulation for Causal Inference from Real-World Data

Vanessa Didelez, **Maria Geers**
Leibniz Institute for Prevention Research and Epidemiology - BIPS, Germany; vdidelez@uni-bremen.de, geers@leibniz-bips.de
Keywords: observational data, avoding self-inflicted biases, comparing treatment strategies

Target trial emulation (TTE) is a general principle to organize and structure the analysis of observational data, such as electronic health records, claims or registry data, so as to minimize common but avoidable sources of bias, e.g. immortal-time bias. Moreover, formulating a target trial is helpful to elicit practically meaningful causal research questions (aka "estimands") with a clear interpretation. The workshop will explain the principle of TTE using examples from cancer screening, drug safety as well as nutritional epidemiology. For instance, we will illustrate how to emulate a target trial on screening colonoscopy, how this avoids design-related and other biases, while showing how results are badly affected if a naive study design is chosen that suffers from these biases. A brief overview of some relevant statistical methods will be given, such as the clone-censor-weight approach or the parametric g-formula. However, as will become clear, TTE is a fundamental principle that can be combined with various causal inference methods.

Organization of the workshop: There will be theoretical parts as well as worked examples, with hands-on tasks for the participants.

Learning outcomes: Participants will (i) be able to recognise avoidable sources of bias in naïve studies using observational data; (ii) become aware of basic techniques to avoid these issues; (iii) acquire a basic understanding of TTE that will facilitate studying the more advanced literature.

Improving Precision and Power in Randomized Trials by Leveraging Baseline Variables

Kelly Van Lancker[1], **Michael Rosenblum**[2], **Josh Betz**[2]
[1]Ghent University, Belgium; [2]Johns Hopkins Bloomberg School of Public Health, Baltimore, U.S.A.; kelly.vanlancker@ugent.be, mrosen@jhu.edu, jbetz@jhu.edu
Keywords: covariate adjustment; causal inference; standardization; treatment policy; robustness; group sequential designs

In May 2021, the U.S. Food and Drug Administration (FDA) released a revised draft guidance for industry on "Adjustment for Covariates in Randomized Clinical Trials for Drugs and Biological Products". Covariate adjustment is a statistical analysis method for improving precision in clinical trials by adjusting for pre-specified, prognostic baseline variables (e.g., age, BMI, comorbidities). The resulting sample size reductions can lead to substantial cost savings, and also more ethical trials since they avoid exposing more participants than necessary to experimental treatments. Though covariate adjustment is recommended by the FDA and the European Medicines Agency, many trials do not fully exploit the available information in baseline variables.

In Part 1, we explain what covariate adjustment is, how it works, when it may be useful, and how to implement it (in a preplanned way that is robust to model misspecification) for a variety of scenarios.

In Part 2, we present a new method that enables us to easily combine covariate adjustment with group sequential designs. This approach can lead to faster trials, without sacrificing validity or power, even when the experimental treatment is ineffective.

In Part 3, we show the impact of covariate adjustment using completed trial datasets in multiple disease areas. We provide step-by-step, clear documentation of how to apply the software in each setting. Participants will have the time to apply software tools on the different datasets.

Model and Algorithm Evaluation in Supervised Machine Learning

Max Westphal, **Rieke Alpers**
Fraunhofer Institute for Digital Medicine MEVIS, Bremen, Germany; max.westphal@mevis.fraunhofer.de, rieke.alpers@mevis.fraunhofer.de
Keywords: prediction; performance; validation; comparison; benchmark

The statistical evaluation of developed models and algorithms is an essential part of applied machine learning and predictive modelling. This half-day course is suitable as a concise introduction or refresher for this important topic. It is divided into three parts with sufficient time for participant questions and breaks in between.

Initially, we will repeat essential machine learning basics and cover core concepts of model evaluation. We will mainly consider classification tasks and the most relevant assessment criteria (discrimination, calibration) but also summarize adaptations for regression and survival problems. In the main part, we discuss common pitfalls (leakage, multiplicity, ….) in model evaluation and appropriate best practices to avoid and/or rectify them. Finally, we touch upon some advanced topics and cover important practical aspects (software, reporting, reproducibility) that are required for a successful evaluation study.

The course contents are illustrated by means of real-world data examples, including R code to showcase how the numerical results were obtained. There are no explicit coding sessions in this short course, so a laptop is not necessarily required. The course materials will be made available so that participants have the opportunity to individually reproduce the numerical examples after the course.

Prerequisites:

- Initial practical experience in applied machine learning
- Basic knowledge of descriptive and inferential statistics

Abstracts for Poster Contributions

Monday, September 4

ID: 134 / Poster 1: 1
Presentation Submissions - Regular Session (Default)
Topic of Submission: Free Contributions
Keywords: Discrete Choice Experiment, design, analysis models, MNL, RPL, HB, Poster ID M1

Statistical approaches to the design and analysis of patient preference studies.

Byron Jones, Maria Costa
Novartis, Switzerland; maria_j.costa@novartis.com

Including patient views into drug development and post-marketing decisions is becoming increasingly important to regulatory agencies, Health Technology Assessment (HTA) bodies and the pharmaceutical industry [1,2,3]. One important contributor to the totality of patient evidence needed by all the aforementioned is a quantitative patient preference study.

This presentation will describe the design and analysis of one major type of patient preference study: the Discrete Choice Experiment (DCE). In a DCE patients are presented with a series of choice tasks, where, in each task, they are presented with profiles of treatments, devices or symptoms and asked to choose the profile they prefer. The basic structure of the design of a DCE is similar to that of a fractional factorial design. Statistical models that are typically used to analyse the preferences obtained in a DCE are, in increasing order of complexity, the multinomial logistic model (MNL), the random parameters logistic model (RPL) and the hierarchical Bayesian model (HB). Optimal designs for a DCE depend on the model chosen and can be obtained using theoretical results or computer search algorithms. A brief description of the different approaches will be given.

The analysis of a DCE will be described using a small case study based on a published DCE [4] that collected preferences from patients suffering with COPD. The results from the different models will be compared and some useful graphical displays of the results will be given.

References

1. https://www.fda.gov/drugs/development-approval-process-drugs/cder-patient-focused-drug-development
2. https://www.ema.europa.eu/en/documents/presentation/presentation-ema/fda-patient-focused-drug-development-ich-reflection-paper-mbonelli-ema_en.pdf
3. Shared decision making NICE guideline Published: 17 June 2021 www.nice.org.uk/guidance/ng197
4. Cook, N.S., Criner, G.J., Burgel, P-R, Mycock, M., Gardener, T., Mellor, P. Hallworth, P., Sully, K., Tatlock, S., Klein, B., Jones, B., Le Rouzic, O., Adams, K., Phillips, K., McKevitt, M. Toyama, K. and Gutzwiller, F. (2022). People living with moderate-to-severe COPD prefer improvement of daily symptoms over the improvement of exacerbations: a multicountry patient preference study. ERJ Open Research, 01 Apr 2022, 8(2):686-2021. DOI: 10.1183/23120541.00686-2021
5. CDER's Patient-Focused Drug Development

ID: 168 / **Poster 1: 2**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), Preclinical drug development, safety and toxicology
Keywords: Design space, desirability, Bayesian modelling, posterior predictive distribution, probability of success, Poster ID M2

From Desirability Towards Bayesian Design Space

Martin Otava
Janssen Pharmaceutical Companies of Johnson & Johnson, Czechia; motava@its.jnj.com

The design space answers basic scientific question: *across the experimental space, which settings lead to high probability of achieving certain quality threshold?* [1] For example in pharmaceutical manufacturing context: the probability that a chemical synthesis will result in yield above 97%.

Traditional setup to answer such question is frequentist modelling of available data from designed experiment and defining a desirability function: function of quality level with its value reflecting our desire for such level [2]. Simplest version for yield example would be a step function that is zero below 97% and is one for any value above 97%. Naturally, argument arises that 96.9% could be also rather acceptable and desirability framework allows us to reflect that by specifying less steep function, e.g. zero below 90% and then linearly increasing from 90 to 97% or even smoother function.

However, there are two main issues with such approach. Firstly, the function is often chosen very arbitrarily in practice as it can be rather unclear which functional shape should be chosen. Moreover, it attempts to precisely quantify the desire of particular quality level, but the evaluation is based solely on average predictions, ignoring any uncertainty in the estimate of the relationship between predictor and response.

The frequentist design space checks whether model predictions at each experimental setting are above pre-specified quality threshold, providing an estimate of probability of interest. However, such estimate is still typically based only on residual error and average prediction. As any simulation from confidence intervals on model parameters is incompatible with frequentist framework, adding uncertainties on various predictors' effect sizes is cumbersome matter.

Bayesian framework allows to address the uncertainty in all parameters directly by conducting posterior simulation [4]. Posterior distribution of response can be used to determine probability of compliance with specifications for a given settings of covariates. Alternatively, the probability itself can be seen as the quantity of interest and calculated from analytical closed-form solution at each iteration to obtain posterior distribution of the probability itself.

In this presentation, we look at the advantages and disadvantages of both desirability and Bayesian design space framework, including the implementation and interpretation challenges. We will demonstrate the flexibility of Bayesian framework and emphasize the need of clearly identifying aforehand what is the ultimate metric of interest. A case study from pharmaceutical manufacturing will be shown (with simulated data), but the applicability of the demonstrated principles is broader: the interpretation of probabilistic statements in Bayesian framework in comparison to counterpart frequentist methods, the correct use of Bayesian results and worthiness of more complicated framework in simple cases.

References:

[1] Harrington Jr. E. C. (1965). The Desirability Function. *Industrial Quality Control:* 21(10): 494-498.

[2] Del Castillo E., Montgomery D. C., McCarville D. R. (1996). Modified Desirability Functions for Multiresponse Optimization. *Journal of Quality Technology,* 28(3): 337-345.

[3] Lebrun P., Boulanger B., Debrus B., Lambert P., Hubert P. (2013) A Bayesian Design Space for Analytical Methods Based on Multivariate Models and Prediction. *Journal of Biopharmaceutical Statistics*, 23(6): 1330-1351

ID: 212 / **Poster 1: 3**
Poster Submission
Topic of Submission: High dimensional data, genetic and x-omics data, Statistical hypothesis testing (covariate adjustment, nonparametric methods, multiple comparisons, …)
Keywords: GWAS, survival, Cox model, testing, genome-wide significance, Poster ID M3

Fast and Precise Survival Testing for Genome-Wide Association Studies

Tong Yu[1], Axel Benner[1], Dominic Edelmann[1,2]
[1]German Cancer Research Center, Heidelberg, Germany; [2]NCT Trial Center, Heidelberg, Germany; dominic.edelmann@dkfz-heidelberg.de

Genome-wide association studies (GWAS) involve testing millions of single-nucleotide polymorphisms for their association with survival responses. To adjust for multiple testing, the genome-wide significance level with $\alpha=5x10^{-8}$ is commonly used. However, standard survival statistics based on the Cox model such as the Wald or the Score test are cannot reliably control the type I error rate for this level. On the other hand, more reliable alternatives, such as the Firth correction, are computationally expensive, making them impractical for large datasets.

We compared the type I error rate, power, and runtime of various Cox model-based survival tests, including the Score test, Wald test, Likelihood ratio test, Firth correction, and a saddle-point approximation based Score test (SPACOX) using simulations and real data from the UK Biobank. Our findings reveal that the Wald and Score tests are highly anti-conservative for low minor allele frequencies (MAFs) and/or event rates, whereas SPACOX is substantially conservative in some settings. Furthermore, these tests exhibit different behavior depending on the direction of the effect. Except for score test-based procedures, the runtime of all tests is prohibitively high, particularly for the Firth correction.

To address this challenge, we propose a fast and precise testing procedure for GWAS based on prescreening via an extremely efficient version of the Score test, followed by testing of the screened subset of genes using the Firth correction or Likelihood ratio test. We demonstrate the performance of our test using simulations and real data from the UK Biobank. Our method provides a practical and accurate alternative for GWAS that can be applied to large datasets.

ID: 214 / **Poster 1: 4**
Poster Submission
Topic of Submission: High dimensional data, genetic and x-omics data
Keywords: proteomics, bipartite graphs, protein quantification, optimization, Poster ID M4

Improved protein quantification by using bipartite peptide-protein graphs

Karin Schork[1,2,3], Michael Turewicz[1,2,4,5], Julian Uszkoreit[1,2,6], Jörg Rahnenführer[3], Martin Eisenacher[1,2]
[1]Medizinisches Proteom-Center, Medical Faculty, Ruhr-University Bochum, Germany; [2]Medical Proteome Analysis, Center for Protein Diagnostics (PRODI), Ruhr-University Bochum, Bochum, Germany; [3]Department of Statistics, TU Dortmund University, Dortmund, Germany; [4]Current address: Institute for Clinical Biochemistry and Pathobiochemistry, German Diabetes Center (DDZ), Leibniz Center for Diabetes Research at the Heinrich Heine University Düsseldorf, Düsseldorf, Germany; [5]Current address: German Center for Diabetes Research (DZD), Partner Düsseldorf, München-Neuherberg, Germany; [6]Current address: Universitätsklinikum Düsseldorf, Düsseldorf, Germany; karin.schork@rub.de

Introduction:

In bottom-up proteomics, proteins are enzymatically digested to peptides (smaller amino acid chains) before measurement with mass spectrometry (MS), often using the enzyme trypsin. Because of this, peptides are identified and quantified directly from the MS measurements. Quantification of proteins from this peptide-level data remains a challenge, especially due to the occurrence of shared peptides, which could originate from multiple different protein sequences.

The relationship between proteins and their corresponding peptides can be represented by bipartite graphs. In this data structure, there are two types of nodes (peptides and proteins). Each edge connects a peptide node with a protein node, if and only if the peptide could originate from a tryptic digestion of the protein. The aim of this study (Schork et al, 2022, PLOS ONE) is to characterize and structure the different types of graphs that occur and to compare them between different data sets. Furthermore, we want to show how this knowledge can aid relative protein quantification. Our focus is especially on gaining quantitative information about proteins with only shared peptides, as they are neglected by many current algorithms.

Methods:

We construct bipartite peptide-protein graphs using quantified peptides from three measured data sets, as well as all theoretically possible peptides from the corresponding protein sequence databases. The structure and characteristics of the occurring graphs are compared between data sets as well as between database (theoretical) and quantitative level.

Additionally, we developed and applied a method that calculates protein ratios from peptide ratios by making use of the bipartite graph structures. For each peptide node, an equation is formed based on the bipartite graph structures and the measured peptide ratios. Protein ratios are estimated by using an optimization method to find solutions with a minimal error term. Special focus lies on the proteins with only shared peptides, which often lead to a range of optimal solutions instead of a point estimate.

Results:

When comparing the graphs from the theoretical peptides to the measured ones, two opposing effects can be observed. On the one hand, the graphs based on measured peptides are on average smaller and less complex compared to graphs using all theoretically possible peptides. On the other hand, the proportion of protein nodes without unique peptides, which are a complicated case for protein quantification, is considerably larger for measured data. Additionally, the proportion of graphs containing at least one protein node with only shared peptides rises, when going from database to quantitative level.

Conclusion:

Large differences between the structures of bipartite peptide-protein graphs have been observed between database and quantitative level as well as between the three analyzed species. In the three analyzed measured data sets, the proportion of protein nodes without unique peptides were 6.3 % (yeast), 46.6 % (mouse) and 55.0 % (human), respectively. Especially for these proteins, the usage of information from the bipartite graph structures for protein quantification is beneficial.

ID: 215 / **Poster 1: 5**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Biomarkers and diagnostics, Prevention and handling of missing data
Keywords: diagnostic study, accuracy, missing values, inconclusive, reference standard, index test, sensitivity and specificity, Poster ID M5

Missing values and inconclusive results in diagnostic studies – a scoping review of methods

Katharina Stahlmann[1], Johannes B. Reitsma[2], Antonia Zapf[1]
[1]Institute of Medical Biometry and Epidemiology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany; [2]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands; k.stahlmann@uke.de
Note: A companion presentation to this contribution will be given at the "68. GMDS-JAHRESTAGUNG 2023".

Introduction:

Inappropriate handling of missing values can lead to biased results in any type of research. In diagnostic accuracy studies, this can have severe consequences by leading to misdiagnosing many patients and assigning them to the wrong treatment. Nonetheless, most diagnostic studies exclude missing values and inconclusive results in either the index test or reference standard from the analysis or apply simple methods resulting in biased accuracy estimates. This may be due to the lack of availability or awareness of appropriate methods. Therefore, the aim of this scoping review was to provide an overview of strategies to handle missing values and inconclusive results in the reference standard or index test in diagnostic accuracy studies.

Methods:

We conducted a systematic literature search in MEDLINE, Cochrane Library, and Web of Science up to April 2022 to identify methodological articles proposing methods for handling missing values and inconclusive results in diagnostic studies. Additionally, reference lists and Google Scholar citations were searched. Besides methodological studies, we also searched for studies applying the proposed methods.

Results:

Of 110 articles included in this review, most (n=67) addressed missing values in the reference standard. A further 15 and 12 articles proposed methods for handling missing values in the index test and in the index and reference test, respectively. Lastly, 17 articles presented methods for handling inconclusive results. Methods for missing values in the index test and inconclusive results encompass imputation, frequentist and Bayesian likelihood, model-based, and latent class methods. Most of these methods can be applied under missing (completely) at random, but only a few also incorporated missing not at random assumptions. While methods for missing values in the reference standard are regularly applied in practice, this is not the case for methods addressing missing values and inconclusive results in the index test.

Discussion:

Missing values and inconclusive results in the index test are commonly not adequately addressed in diagnostic studies despite the availability of various methods based on different assumptions. This may be due in part to the lack of programming code, R packages, or Shiny apps, which would facilitate the application. Our comprehensive overview and description of available methods may be the first step to raising further awareness of these methods and enhancing their application. Nevertheless, future research is needed to compare the performance of these methods under different conditions to give valid and robust recommendations for their usage in various diagnostic accuracy research scenarios. Within our project, we currently work on a simulation study to compare the identified methods in order to make recommendations regarding their application.

ID: 217 / **Poster 1: 6**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), Time-to-Event Analysis, Statistical hypothesis testing (covariate adjustment, nonparametric methods, multiple comparisons, …)
Keywords: Poster ID M6

Evaluating statistical matching methods for treatment effects after different aortic valve replacement surgeries

Veronika Anna Waldorf, Eva Herrmann
Goethe University Frankfurt, Germany; v.waldorf@hotmail.de

For better comparison of treatment effects matching methods are commonly used in observational studies. Besides propensity score matching, regression models and inverse probability weighting schemes are often implemented.

In our simulation study the settings mimic patient groups of the German Aortic Valve Registry (GARY). Those groups represent the two most common aortic valve replacement surgeries: minimal invasive transcatheter aortic valve implantation (TAVI) and surgical aortic valve replacement (SAVR).

Patients presenting with different survival affecting health parameters will receive different treatments suitable to their varying operability. Therefore, patient population within SAVR is mostly in better health condition compared to TAVI. [1]

Even though statistics have shown a supremacy of SAVR surgeries in comparison to TAVI this superiority cannot be taken as fact due to the heterogenous group populations.

In our study setting we performed different propensity score matching methods as well as a weighting scheme and regression model in R using the packages MatchIt, twang and survival [2-4], aiming to decrease the imbalance of both groups and bias in comparing clinical outcome.

The treatment effect was measured with time-to-event analysis. To further answer the usability of these methods we simulated different group sizes from 500 up to 20.000 patients, with 500 simulations in each sample size.

References

1. Hamm CW et al. The German Aortic Valve Registry (GARY): in-hospital outcome. European Heart Journal. 2014; 35: 1588–1598.
2. Ho DE et al. MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. Journal of Statistical Software. 2011; 42: 1–28.
3. Ridgeway G et al. twang: Toolkit for Weighting and Analysis of Nonequivalent Groups. R package version 1.4-9.5. 2016.
4. Therneau T. survival: A Package for Survival Analysis in R. R package version 3.5-3. 2023.

ID: 220 / **Poster 1: 7**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Meta-Analysis and Systematic Reviews
Keywords: meta-analysis, R shiny, user interface, imputation, Poster ID M7

GUiMeta - A new graphical user interface for meta-analyses

Rejane Golbach
Goethe University Frankfurt, Germany; golbach@med.uni-frankfurt.de

The popularity of meta-analyses is rising, which is reflected in the steadily growing number of publications. Many theoretical models are being newly and further developed while at the same time easy-to-learn tools for performing meta-analyses become increasingly important. Several menu-based programs but also R packages are available for conducting meta-analyses. Nevertheless, one of the major difficulties remains the summarization of the data, i.e., performing imputations, which is hardly supported by most programs. The R shiny [1] application GUiMeta provides a solution to this difficulty with a new interface for meta-analyses.

GUiMeta guides users through data entry, data analysis based on state-of-the-art R packages, and interpretation of the results supported by meaningful graphs and statistically substantiated results. One of GUiMeta's major strengths is the adaptive data table, which is provided for documenting effect sizes from systematic reviews and structuring the heterogeneous data from various studies.

Reference:

[1] Chang W, Cheng J, Allaire J, Sievert C, Schloerke B, Xie Y, Allen J, McPherson J, Dipert A, Borges B (2022). _shiny: Web Application Framework for R_. R package version 1.7.2

ID: 240 / **Poster 1: 8**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Clinical trial designs (adaptive and platform designs, external controls, …)
Keywords: Optimus, dose-response, phase 1, cancer, efficacy, Poster ID M8

Impact of dose modifications on the assessment of the dose-tumor size relationship in oncology dose-ranging clinical trials

Francois Mercier[1], **Maia-Iulia Muresan**[2], **Georgios Kazantzidis**[1], **Daniel Sabanes-Bove**[1], **Ulrich Beyer**[1]
[1]F. Hoffmann-La Roche, Switzerland; [2]University of Geneva; francois.mercier@roche.com

Changes in tumor size are among the most regarded endpoints in early clinical trials with solid tumor patients. Indeed, longitudinal data on the sum of lesion diameters (SLD) of target lesions provide insights into the extent and duration of response to treatment. Nonlinear mixed-effect models, referred to as tumor growth inhibition (TGI) models, have proved to accurately capture these data, and they have been used to support dose recommendation and prediction of survival. These models can be fitted with relatively sparse data, and therefore could also be used to inform the dose-response relationship based on phase 1a and/or phase 1b study data. In phase 1 studies, patients are exposed to a range of doses to assess safety/toxicity. However, early insights on efficacy are also instrumental to define the recommended phase 2 dose. In these trials, it is not unusual for participants to omit doses, or to receive doses lower than expected, in order to mitigate toxicity events.

Using simulations, we evaluate the impact of dose modifications on the ability to characterize the dose-response relationship using TGI models. Various scenarios are considered where doses are either reduced or omitted, in trials of sample size ranging from 3 to 10 patients per dose-cohort, and with a proportion of patients impacted by dose modifications ranging from 10% to 50%. In each case, a TGI dose response is fitted to the simulated data. Simulation outputs are expressed in terms of bias and (im)precision of the estimated dose-response model when compared to the hypothetical scenario of no dose modification.

We draw conclusions on the minimum conditions required to study the dose-effect on patients' tumor burden in phase 1 dose-escalation studies.

ID: 247 / **Poster 1: 9**
Poster Submission
Topic of Submission: Clinical trial designs (adaptive and platform designs, external controls, …)
Keywords: analysis of variance, statistical interaction, sample size, Poster ID M9

Simulation based sample size calculation in two-way fixed effects designs including interactions and repeated measures

Louis Rodrigue Macias, Silke Szymczak
Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck,Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck, Germany; louis.macias@uni-luebeck.de

Block designs to compare quantitative outcomes between groups of subjects assigned to a combination of j levels of factor A and k levels of factor B are widely used in animal studies. Existing tools for sample size calculation in this setting require effect size, expressed as partial η2, Cohen's f and f2, among others, to be specified. Jacob Cohen1 provides guidelines on the interpretation of f values with respect to the variability of within group means. These examples have been interpreted as rules of thumb for small, medium and large effect sizes. However, determining the f values corresponding to expected group means in a j by k design is not intuitive. This is especially the case when there is interaction in the factors' effects. Moreover, when one of the factors is a repeated measure, within subject correlation must be considered. Finally, sample size estimation for non parametric tests are not easily implemented. For these reasons, we have implemented several R functions that perform these tasks by simulating the planned experimental setting for both independent and repeated measurements.

Sample size required for a power 1-β with a type 1 error α is calculated by estimating the proportion of iterations in which the p value of interest, either effect for factor A, factor B or their interaction, is smaller than α. The simulation input is a j by k matrix of expected means, a matrix of standard deviations (SD) with the same dimensions and group sample size which may be a single integer for the case of a balanced design or a matrix, if the design is unbalanced. A helper function is available to create the means and SD matrices. Input for this function includes a reference mean, which would typically correspond to the expected mean in group j,k =1, expected change from reference mean by each factor level in the multiplicative scale and expected deviation from linear effects because of interaction.

Sample sizes requirements were compared to those obtained for ANOVA by G*Power2. However, our implementation has the advantage of allowing unbalanced designs and to also estimate power for non-parametric tests. Computation time is < 5 minutes for 1000 iterations in a 2 by 3 design on a 3.9 GHz processor in the independent measurements case and ~2 hours in a 6 by2 repeated measurements design, depending on the sample size space explored. We plan to provide these functions as an R package that can be part of a series of useful tools when planning a two-way fixed effects study.

References:

1. Cohen. J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
2. Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods, 39, 175-191.

ID: 259 / **Poster 1: 10**
Poster Submission
Topic of Submission: Preclinical drug development, safety and toxicology
Keywords: experimental design, reproductive toxicology, Poster ID M10

A slight modification of the experimental design can cause a substantial knowledge gain in non-clinical studies with female rats

Monika Brüning[1], **Bernd Baier**[2], **Bernd-Wolfgang Igl**[1]
[1]Global Biostatistics and Data Sciences, Boehringer Ingelheim Pharma GmbH & Co. KG, Germany; [2]Nonclinical Drug Safety, Boehringer Ingelheim Pharma GmbH & Co. KG, Germany; monika.bruening@boehringer-ingelheim.com, bernd-wolfgang.igl@boehringer-ingelheim.com

In pharmaceutical industry, various non-clinical studies are performed to secure the safety of potential drug candidates for use in humans. To address potential effects on fertility of patients so called “Fertility and Early Embryonic Developmental” (FEED) animal studies are conducted. Therein, the detection of possible effects of the test item on regular estrous cyclicity of about 4 days in female rats is a relevant endpoint, which is usually measured by vaginal swaps and cytological analyses of the estrous stage before and during treatment.

As estrous cyclicity causes substantial changes in the general activity level of female rats, we compared the estrous cycle data with body weight changes. Body weight is a highly relevant and informative endpoint in toxicological *in vivo* studies usually taken at least to times per week in rodents. To increase the time resolution during the estrous cycle we took daily body weights.

We now have analyzed the possibility to estimate relevant information on the female reproductive status by simply weighing the animals daily. This has several positive consequences for the conducting laboratory and is also of concern from an animal welfare perspective. In addition, it might not only be relevant in dedicated FEED studies but could be applied more generally in studies on female rats: with simple body weight measurements the complex hormonal and behavioral changes could be monitored allowing for a better interpretation of data on an individual and population level.

ID: 264 / **Poster 1: 11**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Real world data and evidence, Software Engineering, Preclinical drug development, safety and toxicology
Keywords: Toxicology, bio-assay, bootstrap calibration, assay validation and qualification, pre-clinical statistics, Poster ID M11

Prediction intervals for overdispersed Poisson data and their application to historical controls

Max Menssen
Leibniz Universität Hannover, Germany; menssen@cell.uni-hannover.de

Toxicological studies are a class of biological trials which are aimed to evaluate the toxicological properties of chemical compounds on model organisms. Typical toxicological studies are comprised of an untreated control group and several cohorts that were treated with the compound of interest.

If the same type of study is run for several times, using the same model organism and the same experimental setup, the knowledge about the baseline reaction obtained from the (historical) untreated control groups is rising with every new trial. Hence, most guidelines from the OECD require the verification of the actual control group based on historical control data. This can be done by the application of prediction intervals.

In several study-types such as the Ames assay, the endpoint of interest is comprised of counted observations (e.g. number of reverant bacteria colonies per petridish). In this case, the historical control data can be modelled to be overdispersed Poisson. Hence, an asymptotic prediction interval that allows for overdispersion is proposed. Furthermore, it will be demonstrated how to use bootstrap calibration in order to enhance the intervals small sample properties.

The proposed methodology is implemented in the R package predint and its application will be demonstrated based on a real life data set.

ID: 265 / **Poster 1: 12**
Poster Submission
Topic of Submission: Statistical hypothesis testing (covariate adjustment, nonparametric methods, multiple comparisons, …)
Keywords: Simultaneous confidence band, Confidence sets, Minimum area; Linear regression, Percentile line, Poster ID M12

Minimum Volume Confidence Set Optimality for Simultaneous Confidence Bands for Percentiles in Linear Regression

Lingjiao Wang[1], **Yang Han**[1], **Wei Liu**[2], **Frank Bretz**[3]
[1]Department of Mathematics, University of Manchester, UK; [2]School of Mathematical Sciences & Southampton Statistical Sciences Research Institute, University of Southampton, UK; [3]Novartis Pharma AG, Basel, Switzerland;
lingjiao.wang@postgrad.manchester.ac.uk

Simultaneous confidence bands for a percentile line in linear regression have been considered by several authors and the average width of a simultaneous confidence band has been widely used as a criterion for the comparison of different confidence bands. In this work, exact symmetric and asymmetric simultaneous confidence bands over finite covariate intervals are considered, and the area of the confidence set that corresponds to a confidence band is used as the criterion for the comparison. The optimal simultaneous confidence band is found under the minimum area confidence set (MACS) or minimum volume confidence set (MVCS) criterion. The area of corresponding confidence sets for asymmetric simultaneous confidence bands is uniformly and can be very substantially smaller than that for the corresponding exact symmetric simultaneous confidence bands. Therefore, asymmetric simultaneous confidence bands should always be used under the MACS criterion. A real data example is included for illustration.

ID: 271 / **Poster 1: 13**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Biomarkers and diagnostics
Keywords: Agreement, limits of agreement, tolerance limits, differential bias, proportional bias, method comparison., Poster ID M13

Use of clinical tolerance limits for assessing agreement

Patrick Taffé
University of Lausanne (UNIL), Switzerland; patrick.taffe@unisante.ch

Bland & Altman's limits of agreement (LoA) is one of the most used statistical methods to assess the agreement between two measurement methods (for example between two biomarkers). This methodology, however, does not directly assess the level of agreement and it is up to the investigator to decide whether or not disagreement is too high for the two methods to be deemed to be interchangeable.

To more directly quantify the level of agreement, Lin et al. (2002) have proposed the concept of « coverage probability », which is the probability that the absolute difference between the two measurements made on the same subject is less than a pre-defined level. Their methodology, however, does not take into account that the level of agreement might depend on the value of the true latent trait. In addition, homoscedastic measurement errors are implicitly assumed (an often too strong assumption) and the presence of a possible bias is not assessed. For these reasons, Stevens et al. (2017, 2018) have extended this methodology to allow the coverage probability to depend on the value of the latent trait, as well as on the amount of bias, and called their extended agreement concept « probability of agreement ».

In this study, we have further extended this methodology by relaxing the strong parametric assumptions regarding the distribution of the latent trait and developing inference methods allowing to compute both pointwise and simultaneous confidence bands. Our methodology requires repeated measurements for at least one of the two measurement methods and accommodates heteroscedastic measurement errors. It performs often very well even with only one measurement for one of the two measurement methods and at least 5 repeated measurements from the other. It circumvents some of the deficiencies of LoA and provides a more direct assessment of the agreement level.

References

1. Lin L, Hedayat AS, Sinha B, and Yang M. Statistical methods in assessing agreement: models, issues, and tools. *JASA* 2002; **97**: 257-270.
2. Stevens NT, Steiner SH, and MacKay RJ. Assessing agreement between two measurement systems: an alternative to the limits of agreement approach. *Stat Meth Med Res* 2017; **26**: 2487-2504.
3. Stevens NT, Steiner SH, and MacKay RJ. Comparing heteroscedastic measurement systems with the probability of agreement. *Stat Meth Med Res* 2018; **27**: 3420-3435.
4. Taffé P. Use of clinical tolerance limits for assessing agreement. *Stat Meth Med Res* 2023; **32**: 195-206.

ID: 277 / **Poster 1: 14**
Poster Submission
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), High dimensional data, genetic and x-omics data
Keywords: Coefficient of determination; Explained variation; GWAS; Hierarchical data, Relevance ranking, Linear Mixed Model, Variance decomposition, Poster ID M14

Decomposition of Explained Variation in the Linear Mixed Model

Nicholas Schreck, **Manuel Wiesenfarth**
DKFZ Heidelberg, Germany; nicholas.schreck@dkfz-heidelberg.de

The concepts of variance decomposition and explained variation are the basis of relevance assessments of factors in ANOVA, and lead to the definition of the widely applied coefficient of determination in the linear model. In the linear mixed model, the assessment and comparison of the dispersion relevance of explanatory variables associated with fixed and random effects still remains an important open practical problem. To fill this gap, our contribution is two-fold. Firstly, based on the restricted maximum likelihood equations in the variance components form of the linear mixed model, we prove a proper decomposition of the sum of squares of the dependent variable into unbiased estimators of interpretable estimands of explained variation. Our result leads us to propose a natural extension of the well-known adjusted coefficient of determination to the linear mixed model. Secondly, we allocate the novel unbiased estimators of explained variation to specific contributions of covariates associated with fixed and random effects within a single model fit. These parameter-wise explained variations constitute easily interpretable quantities, assessing dispersion relevance of covariates associated with both fixed and random effects on a common scale, and thus allowing for a covariate ranking. Our approach is made readily available in the user-friendly R-package ``explainedVariation'' and its usefulness is illustrated in public datasets.

ID: 281 / **Poster 1: 15**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Machine Learning and Data Science, Meta-Analysis and Systematic Reviews
Keywords: text classification, machine learning, abstract screening, systemtic review, Poster ID M15

Text classification to automate abstract screening using machine learning

Johannes A. Vey[1], **Samuel Zimmermann**[1], **Maximilian Pilz**[1,2]
[1]Institute of Medical Biometry, University of Heidelberg, Germany; [2]Department Optimization, Fraunhofer Institute for Industrial Mathematics (ITWM), Germany; vey@imbi.uni-heidelberg.de
Note: A companion presentation to this contribution will be given at the "68. GMDS-JAHRESTAGUNG 2023".

Systematic reviews synthesize all available evidence on a specific research question. A paramount task in this is the comprehensive literature search, which should be as extensive as possible to identify all relevant studies and reduce the risk of reporting bias. The identified studies need to be screened according to defined inclusion criteria to address the research question. As a consequence, screening the identified studies is time-consuming, resource intensive and tedious for all researchers involved. In the first stage of this process, the title-abstract screening (TIAB), abstracts of all initially identified studies are screened and classified regarding their inclusion or exclusion for full-text screening. Conventionally, this is accomplished by two independent human reviewers. In the last years, there has been some research to automate the literature search and screening processes [1-2].

We present a semi-automated approach to TIAB screening using natural language processing (NLP) and machine learning (ML) based classification that was applied within a systematic review project on the reduction of surgical site infection incidence in elective colorectal resections.

The total 4460 identified abstracts were randomly split into a training (1/3) and test set (2/3). The titles and abstracts of the publications were processed by methods of NLP to transform the plain language into numerical matrices. Based on the processed training data, variable selection conducting Elastic Net regularized regression was performed. Subsequently, different ML algorithms (Elastic Net, Support Vector Machine, Random Forest, and Light Gradient Boosting Machine) were trained using 5-fold cross-validation and grid search for the respective tuning parameters. The AUC value was used as an optimization criterion and the decision of the two human reviewers was used as the reference. The algorithms were evaluated on the test set.

The Random Forest showed the highest performance in the test set (AUC: 96%). Choosing a cut-off to avoid missing any relevant abstract (n=136) resulted in only 755 false positives (FP rate: 26.5%). Conversely, 2089 abstracts were correctly classified as to be excluded (FN rate: 0%). Further investigations were done on the minimal number of abstracts needed to validly train the Random Forest. In our case study, the manual TIAB screening workload for the second reviewer could be reduced by about 70%.

We propose an approach where a ML model can replace one human reviewer after being trained on a sufficient number of abstracts. The second reviewer only needs to get involved in cases of discrepancies between the decision of the first reviewer and the classification model.

References:

1. Our ML-based text classification approach proved to be powerful, adaptable, and it considerably reduced human workload in creating systematic reviews.
2. O'Mara-Eves, A., Thomas, J., McNaught, J. et al. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Syst Rev* 4, 5 (2015).
3. Lange, T, Schwarzer, G, Datzmann, T, Binder, H. Machine learning for identifying relevant publications in updates of systematic reviews of diagnostic test studies. *Res Syn Meth*. 2021; 12: 506– 515.

ID: 289 / **Poster 1: 16**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Epidemiology, Meta-Analysis and Systematic Reviews, Real world data and evidence
Keywords: truncated normal, mixed logistic regression, systematic review, Poster ID M16

New formulae for the meta analysis of odds ratios from observational studies with classed continuous exposure

Reinhard Vonthein
Universität zu Lübeck, Germany; reinhard.vonthein@uni-luebeck.de

Observational studies with continuous exposure often report dichotomous outcomes by exposure class, where exposure classes are delimited by quantiles, e.g. quartiles. The obvious meta analysis, namely mixed logistic regression, will rely on exposure values that are representative of these classes. Such representatives could be chosen as middle of the class in terms of exposure or of probability, like uneven octiles which require the assumption of a distribution. They could be estimated means of that distribution truncated at the quantiles or, as a partly distribution-free method, be calculated by the “trapezoidal rule”, i.e. as expectation of a truncated distribution with a straight line connecting the normal density at the quantiles. First, some formulae are presented to estimate moments of the latent exposure distribution from quantiles complementing those formulae published before (Wan et al. 2014). Then, formulae are given for the method of moments estimation of means of truncated distributions. After that, the different options to choose representative values are discussed. Finally, they are compared in a simulation study. Scenarios of the simulation study are inspired by a real application with OR of 1.4 estimated from 12 articles, the practical problems of which will beare presented at the GMDS2023. Although some articles reported clearly lognormal exposure, misspecification is included, so that the merit of the partly distribution-free method becomes clearer. Only the trapezoidal rule and the estimation of means of truncated distributions give representatives that are confined to the class they should represent. When data were generated according to the assumed distribution, the distribution-free method had a higher standard error, but was more robust under misspecification.

Reference:

Wan X, Wang W, Liu J, Tong T. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. Medical Research Methodology 2014, 14:135 http://www.biomedcentral.com/1471-2288/14/135

ID: 296 / Poster 1: 17
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), Epidemiology
Keywords: binary logistic regression, biologic interaction, epidemiology, public health, stochastic dependence, Poster ID M17

How to statistically model biologic interactions

Carolin Malsch
University of Greifswald, Germany; carolin.malsch@uni-greifswald.de

Even after decades of statistical modeling, there is still no clear concept how to assess interaction effects of a set of binary factors on a binary response variable. The reason for this seems to be a lack of clarity about how absence of biologic interaction is modeled.

Biological interaction between two risk factors is often understood as either a deviation from additivity of the absolute effects of two (or more) factors, or non-zero coefficients for interaction terms in binary logistic regression. Both approaches are incorrect.

The mathematically adequate concept for modeling biological (non-)interaction in the given context is stochastic (in-)dependence. Hence, strategies and software recommendations provided in the literature to date are misleading and need correction.

Affected by this misunderstanding is also how logistic regression analysis, the most common approach to model the joint effect of two or more factors on a binary response variable in health research, is conducted in application. In most cases, only main effects are estimated in the regression function while interaction terms are omitted completely. Only sometimes a selection of interaction terms is taken into account with the aim to assess biologic interaction.

In the binary logistic regression model, interaction terms do not reflect biological interactions in general. For example, they are inevitably needed to model stochastic, and thus biologic, independence. On the other hand, coefficients of interaction terms take on value zero when a special type of stochastic dependence is present. More precisely, when the conditional probability under presence of multiple factors is described properly by the logistic regression function including only main effects. This form of dependency is always of synergistic character. However, such a strong assumption is certainly valid in the rarest of cases.

Missing out on interaction terms in the logistic regression model leads to severely biased estimates and easily causes misleading interpretation. This is particularly worrying given that results from studies in epidemiology, health services and public health eventually affect clinical and public health recommendations.

To resolve these problems, this contribution seeks to clarify (a) how biologic interactions in a data set are correctly assessed using stochastic (in-)dependence, (b) why interaction terms in a binary regression model must be considered in the regression function and (c) which value they take on in case of absence of biologic interaction.

The related theory is presented and demonstrated on examples. Further, other approaches to assess biologic interactions from data are critically discussed.

ID: 304 / **Poster 1: 18**
Poster Submission
Topic of Submission: Clinical trial designs (adaptive and platform designs, external controls, …)
Keywords: allocation ratio; cost efficiency; biosimilar trials; sample size, Poster ID M18

Alternative allocation ratios leading to more cost-efficient biosimilar trial designs

Natalia Krivtsova, Rachid El Galta, Jessie Wang, Arne Ring
Hexal AG, Germany; natalia.krivtsova@sandoz.com

Objective

Balanced randomization ratios are standard in clinical trials (e.g. in a parallel 2-group design balanced randomization leads to lowest total sample size), but not mandatory. The costs of some biologic treatments used as comparators in biosimilar trials are very high. Exploring different allocation ratios could lead to lower total trial costs while maintaining the same study power.

Methods

Good cost estimates for drug products and study conduct are required to initiate the calculation. We are separating the per-patient study conduct costs ($Cost_{Conduct}$) from the non-patient related costs ($Cost_{Non\text{-}patient}$) with total trial cost for Phase III trial calculated as:

$$Cost_{total}=N_{Test}*Cost_{Test}+N_{Ref}*Cost_{Ref}+(N_{Test}+N_{Ref})*Cost_{Conduct}+Cost_{Non\text{-}patient}$$

When the costs of two drug treatment ($Cost_{Test}$ and $Cost_{Ref}$) are very different, the patient numbers N_{Test} and N_{Ref} could be adapted while maintaining the same study power. The objective of this research is to identify the optimal allocation ratio $R=N_{Test}/N_{Ref}$ that leads to the smallest total cost.

In biosimilar trials, the equivalence testing is performed using the two one-sided tests (TOST) procedure. The primary endpoint is efficacy endpoint (i.e. response rate) for Phase III trial and PK endpoint (i.e. AUC_{inf}) for Phase I trial, respectively, in the example below.

For Phase I, where there are two reference drugs and two allocation ratios, optimal allocation ratios for 3-way similarity were calculated in R by:

- determining the required sample sizes for a range of allocation ratios using simulations,
- evaluating the cost function for each and
- finding the allocation ratios yielding the lowest cost.

Results

The method is illustrated using fictive, but realistic numerical examples of Phase I and Phase III trials.

Phase III trial assumptions included: power of 90%, equivalence margin of 10% with reference arm expected response rate of 70%, $Cost_{Test}$=5kEUR, $Cost_{Ref}$=50kEUR, $Cost_{Conduct}$=40kEUR and $Cost_{Non\text{-}patient}$=25mEUR.

In a balanced design 1214 subjects were required (leading to 107mEUR total cost) and optimal R=1.4 in cost-optimized design led to total trial cost reduction of 2.34mEUR. More practical R=1.5 (3:2 allocation ratio) required 1265 (759, 506) subjects (4% increase) and led to total trial cost reduction of 2.25mEUR (2.1%).

Phase I trial assumptions included: power of 90%, CV of 35%, $Cost_{Test}$=3kEUR, $Cost_{RefEU}$=30kEUR, $Cost_{RefUS}$=60kEUR, $Cost_{Conduct}$=20kEUR and $Cost_{Non\text{-}patient}$=20mEUR.

In a balanced design 279 subjects were required (29.2mEUR total cost) and in cost-optimized design with 3:2:2 randomization ratio – 280 (120, 80, 80), with just one subject more and a total trial cost reduction of 1.07mEUR (3.7%).

Discussion

A fixed optimized randomization ratio does not require any adjustment of analysis methods. It is advantageous in terms of increasing own safety data base (as number of patients on biosimilar drug is larger). There will be additional considerations regarding the practicality of the implementation and the acceptance of the patients. For Phase III it will also mean slightly prolonged trial duration (due to 4% increase in total sample size).

ID: 305 / **Poster 1: 19**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Epidemiology, Real world data and evidence
Keywords: Air pollution, environmental stressors, meteorological data, cross correlation, LISA, Poster ID M19

Exploring the spatiotemporal relationship between air pollution and meteorological conditions in Baden-Württemberg (Germany)

Leona Hoffmann[1], Lorenza Gilardi[2], Marie-Therese Schmitz[3], Thilo Ebertseder[2], Michael Bittner[2], Sabine Wüst[2], Matthias Schmid[3], Jörn Rittweger[1,4]

[1]Institute of Aerospace Medicine, German Aerospace Center (DLR), Cologne, Germany; [2]German Remote Sensing Data Center, German Aerospace Center (DLR), Oberpfaffenhofen, Germany; [3]Institut of Medical Biometry, Informatics and Epidemiology, University Hospital Bonn, Bonn, Germany; [4]Department of Pediatrics and Adolescent Medicine, University Hospital Cologne, Cologne, Germany; leona.hoffmann@dlr.de

Background: In the Epidemiological analysis of health data as the response variable for environmental stressors, a key question is to understand the interdependencies between environmental variables and which variables to include in the statistical model. Various meteorological (temperature, ultraviolet radiation, precipitation, and vapor pressure) and air pollution variables (O3, NO2, PM2.5, and PM10) are available at the daily level for Baden-Württemberg (Germany). This federal state covers both urban and rural areas.

Methods: A spatial and temporal analysis of the internal relationships is performed using a) cross-correlations, both on the grand ensemble of data as well as with subsets, and b) the Local Indications of Spatial Association (LISA).

Results: Meteorological and air pollution variables are strongly correlated between and among themselves, with specific seasonal and spatial features. For example, Nitrogen dioxide and Ozone are strongly interdependent, and the Pearson correlation coefficient varies with time. In January, there is a negative correlation of -0.84 whereas in April, the correlation coefficient is -0.47, in July 0.45, and in October -0.54. For Ozone and Nitrogen dioxide, a shift of the correlation direction as a function of temperature and UV radiation can be observed, confirmed by cross-correlation. Spatially, NO2, PM2.5, and PM10 concentrations are significantly higher in urban than rural regions. For O3, this effect is reversed. As confirmed also by LISA analysis, where distinct hot and cold spots of the different environmental stressors could be identified. In addition, a linear regression analysis suggests that PM10 variation is almost entirely explained by PM2.5 and vapor pressure by temperature.

Conclusion: The results found are generally compatible with the expected dependencies. Thus, our investigation demonstrates that there are variables with similar temporal and spatial characteristics that should be adequately addressed in analyses of health and environmental stressors. Simplification strategies could e.g. discard redundant variables such as PM10 when PM2.5 is available. However, a reduction to one single variable is not helpful due to the complex relationships between meteorological and air pollution variables.

ID: 307 / **Poster 1: 20**
Poster Submission
Topic of Submission: Clinical trial designs (adaptive and platform designs, external controls, …), Statistical modelling (regression modelling, prediction models, …), Software Engineering
Keywords: R package, Bayesian dynamic borrowing, MCMC, Simulation study, Poster ID M20

Psborrow: an R package for complex, innovative trial designs using Bayesian dynamic borrowing

Matthew Secrest[2], Isaac Gravestock[1], Jiawen Zhu[2], Herb Pang[2], Daniel Sabanes Bove[1]
[1]F. Hoffman-La Roche, Switzerland; [2]Genentech, South San Francisco, CA, USA; isaac.gravestock@roche.com

While the randomized controlled trial (RCT) comparing experimental and control arms remains the gold standard for evaluating the efficacy of a novel therapy, one may want to leverage relevant existing external control data to inform the study outcome. For instance, in certain indications such as rare diseases, it may be difficult to enroll sufficient patients to adequately power an RCT. External control data can help increase study power in these situations. External control data can also benefit trials by shortening trial duration or enabling more subjects to receive the experimental therapy. However, analysis of external control data can also introduce bias in the event that the RCT control arm and external control arm are incomparable (e.g., because of confounding, different standards of care, etc). One method for incorporating external control data that mitigates bias is Bayesian dynamic borrowing (BDB). In BDB, information from the external control arm is borrowed to the extent that the external and RCT control arms have similar outcomes.

The implementation of BDB is computationally involved and typically requires Markov chain Monte Carlo (MCMC) sampling methods. To overcome these technical barriers and accelerate the adoption of BDB, we developed the open-source R package '**psborrow2**'. The package has two main goals: First, '**psborrow2**' provides a user-friendly interface for analyzing data with BDB without the need for the user to compile an MCMC sampler. Second, 'psborrow2' provides a framework for conducting simulation studies to evaluate the impact of different trial and BDB parameters (e.g., sample size, covariates) on study power, type I error, and other operating characteristics.

'**psborrow2**' is an open-source package hosted on GitHub (github.com/Genentech/psborrow2) which is freely available for public use. One important focus of 'psborrow2' development has been modular functions and classes to simplify the user experience and, importantly, promote collaboration with the broader statistical community. New methods are easily incorporated into the package, and users are encouraged to consider contributing. We made the package accessible to people who are working in clinical trial design or analysis who have some familiarity with hybrid control concepts.

ID: 310 / **Poster 1: 21**
Poster Submission
Topic of Submission: Real world data and evidence
Keywords: Implementation science, evidence-based practice, study design, control group, Poster ID M21

Implementation science what, why, how and some statistical challenges

Nathalie Barbier, David Lawrence, Cormelia Dunger-Baldauf, Andrew Bean
Novartis, Switzerland; nathalie.barbier@novartis.com

Interventions and evidence-based practices that are poorly implemented often do not produce expected health benefits.

Implementation science (IS) is the scientific study of methods and strategies that facilitate the uptake of evidence-based practice and research into regular use by practitioners and policymakers. The goal of IS is not to establish the health impact of a clinical innovation e.g. a new treatment, but rather to identify and address the factors that affect its uptake into routine use.

Different non-standard study designs are often used to provide evidence on the impact of strategies to improve uptake. This poster will outline some of these designs and the statistical challenges they may entail, for example, if no control group is used, causal inference issues or challenges with missing values.

ID: 322 / **Poster 1: 23**
Poster Submission
Topic of Submission: Estimands and causal inference, Preclinical drug development, safety and toxicology
Keywords: causal inference, preclinical research, bias, directed acyclic graphs, Poster ID M23

Directed acyclic graphs for preclinical research

Collazo Anja
Berlin Institute of Health, Germany; anja.collazo@bih-charite.de

A central goal of preclinical trials is to reduce the uncertainty about the direction and magnitude of an effect, which is causally attributed to an exposure or intervention of interest. Yet, without high internal validity in estimating outcomes, the experimental results remain either scientifically inconclusive - wasting laboratory animals and resources – or lead to erroneous decisions-making potentially harming patients. Thus, experiments with low scientific rigor raise major ethical concerns. Directed acyclic graphs are visual representations of simplified, key components of a hypothesized causal structure widely used in epidemiological observational research to clarify types and mechanism of biases. Here, we introduce directed acyclic graphs as a tool to improve bias detection in preclinical experiments and argue that causal models can help scientific communication and transparency in preclinical research. We present examples for biased effect estimates from preclinical research expressed in DAGs such as confounding and collider stratification. We show how DAGs can inform the choice of study design, the selection of the smallest subset of variables for measurement and guide the analysis strategy.

Tuesday, September 5

ID: 325 / Poster 2: 1
Poster Submission
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), High dimensional data, genetic and x-omics data, Time-to-Event Analysis
Keywords: Multi-state models, penalized Cox regression, homogeneous effects, sparse group fused lasso, high-dimensional molecular data, Poster ID T1

Model selection strategies for penalized multi-state models incorporating molecular data

Kaya Miah, Annette Kopp-Schneider, Axel Benner
Division of Biostatistics, German Cancer Research Center (DKFZ), Heidelberg, Germany; k.miah@dkfz.de

In medical research, common prediction models still predominantly make use of composite endpoints such as progression- or event-free survival. However, these time-to-first-event outcomes do not incorporate important aspects of the individual disease pathway and therapy sequences. In the era of precision medicine with increasing molecular information, the use of a multi-state model is essential to more accurately capture pathogenic disease processes and underlying etiologies.

Especially the availability of big data with a large number of covariates presents several statistical challenges for model building. Effective data-driven model selection strategies for multi-state models are essential to determine an optimal, ideally parsimonious model based on high-dimensional data. Established methods incorporate regularization in the fitting process in order to perform variable selection. A useful technique to reduce model complexity is to combine homogeneous covariate effects for distinct transitions based on a reparametrized model formulation. We integrate this approach to data-driven variable selection by extended regularization methods for model selection within multi-state model building. We propose the sparse group fused lasso penalized Cox-type regression in the framework of multi-state models combining the penalization concepts of pairwise differences of covariate effects along with transition grouping.

This raises the following challenges: First, multiple heterogeneous transitions have to be considered for consecutive treatment phases within the multi-state model. Furthermore, the number of transitions with fewer observations increases during the course of the sequential event history. Finally, model selection procedures have to be efficiently implemented in large-scale multi-state settings.

Thus, model selection strategies for multi-state endpoints are substantial for a more precise understanding and interpretation of individual disease pathways, specific oncological entities along with their precision therapies as well as improved personalized prognoses.

ID: 399 / **Poster 2: 12**
Poster Submission
Topic of Submission: High dimensional data, genetic and x-omics data
Keywords: genetic linkage analysis, microbiome, IBD, Poster ID T12

Network-based quantitative trait linkage analysis of microbiome composition in inflammatory bowel disease families

Arunabh Sharma, Olaf Junge, Silke Szymczak, Malte Rühlemann, Janna Enderle, Stefan Schreiber, Matthias Laudes, Andre Franke, Wolfgang Lieb, Michael Krawczak, Astrid Dempfle
University of Kiel, Germany; sharma@medinfo.uni-kiel.de

Introduction: Inflammatory bowel disease (IBD) is characterized by a dysbiosis of the gut microbiome that results from the interaction of the constituting taxa with one another, and with the host. At the same time, host genetic variation is associated with both IBD risk and microbiome composition.

Methods: In the present study, we defined quantitative traits (QTs) from modules identified in microbial co-occurrence networks to measure the inter-individual consistency of microbial abundance and subjected these QTs to a genome-wide quantitative trait locus (QTL) linkage analysis.

Results: Four microbial network modules were consistently identified in two cohorts of healthy individuals, but three of the corresponding QTs differed significantly between IBD patients and unaffected individuals. The QTL linkage analysis was performed in a sub-sample of the Kiel IBD family cohort (IBD-KC), an ongoing study of 256 German families comprising 455 IBD patients and 575 first- and second-degree, non-affected relatives. The analysis revealed five chromosomal regions linked to one of three microbial module QTs, namely on chromosomes 3 (spanning 10.79 cM) and 11 (6.69 cM) for the first module, chr9 (0.13 cM) and chr16 (1.20 cM) for the second module, and chr13 (19.98 cM) for the third module. None of these loci have been implicated in a microbial phenotype before.

Discussion: Our study illustrates the benefit of combining network and family-based linkage analysis to identify novel genetic drivers of microbiome composition in a specific disease context.

ID: 350 / **Poster 2: 3**
Poster Submission
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), Time-to-Event Analysis
Keywords: Poster ID T3

Case-Only Analysis in Prospective Cohort and Case-Cohort Studies with Time-to-Event Endpoints

Sandra Freitag-Wolf, Oluwabukunmi Mercy Akinloye, Astrid Dempfle
Institute of Medical Informatics and Statistics, Kiel University, Germany; freitag@medinfo.uni-kiel.de

We present a modification of the case-only (CO) approach to time-to-event data for testing multiplicative interactions between binary risk factors from prospective cohort and case-cohort studies. Motivated by a real data example of a cohort study on survival after cardiovascular surgery, we use the event time information to select only patients who died early (i.e. before a pre-specified time point) and modified the CO approach to time-to-event data. The relevant key assumptions of the CO design, rare events and independence between the factors in the general population, were fulfilled. In a simulation study, we investigated the CO approach in the cohort and case-cohort design with time-to-event outcome and compared results from both designs to the classical Cox proportional hazard and logistic regression (LR). For the LR approach, the same cases as in the CO approach were used and censored observations were considered as 'controls' in a restricted follow-up scheme in the cohort design and a random subsample in the case-cohort design. In our conducted scenarios with varied event rates and main effects, the applied CO approach was consistently valid in the cohort settings and had a similar power as the alternative analyses. In the case-cohort design, the CO approach was distinctly more powerful than standard LR or Cox regression but in the presence of main effects the estimators are biased and consequently the type I error rate slightly inflated. In summary, under a variety of specific circumstances, the CO approach is as powerful for time-to-event data as the standard models in the cohort framework and even more powerful in the case-cohort framework.

ID: 357 / **Poster 2: 4**
Poster Submission
Topic of Submission: Prevention and handling of missing data, Personalized health care
Keywords: Simulation study, missing predictor data, prediction model, Poster ID T4

Is the performance of a prediction model affected by the way of imputing missing predictor data? A simulation study

Manja Deforth[1], **Georg Heinze**[2], **Ulrike Held**[1]
[1]Department of Biostatistics at the Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Zurich, Switzerland; [2]Center for Medical Data Science, Institute of Clinical Biometrics, Medical University of Vienna, Vienna, Austria;
manjaelisabeth.deforth@uzh.ch

Validated prediction models to estimate the patients' risk of developing a particular disease or a condition in the future can be a useful decision tool for individualized treatment decisions. Prediction models are usually developed and validated based on data from observational studies, where missing values resulting from different underlying missingness generating mechanisms are a typical problem. In a simulation study conducted by Marshall et al. [2010], the influence of different imputation methods in a Cox proportional hazards framework were investigated. The authors emphasized that further simulation studies in different clinical contexts and underlying distributions of the predictors are required to assess the generalizability of the results. In the clinical context of long COVID, we study the influence of commonly used imputation methods, including multivariate imputation by chained equations (mice), multiple imputation using additive regression, bootstrapping, and predictive mean matching (aregImpute), and non-parametric missing value imputation using random forest methodology (missForest), to handle missing predictor data on model performance. For the set-up of the simulation study, the recommendations of Morris et al. [2019] and Burton et al. [2006] are followed. The data underlying the simulation study is generated based on a Swiss multicenter prospective cohort study [Deforth et al., 2022] to reproduce a "real-world setting". Several scenarios with different underlying data missingness generating mechanisms, percentage of missing values, stronger and weaker regression coefficients, and varying sample sizes are investigated in a traditional statistical framework. Excluding observations with missing data, complete case, is set as a benchmark. Model performance is assessed in external validation data (without missing values) based on model discrimination ability, calibration-in-the-large, calibration slope and scaled Brier score. The results of the simulation study will be presented in the framework of a neutral comparison study.

References

1. A Marshall, D G Altman, P Royston, and R L Holder. Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. BMC Medical Research Methodology, 10(1):7, 2010. doi: 10.1186/1471-2288-10-7.
2. T P Morris, I R White, and M J Crowther. Using simulation studies to evaluate statistical methods. Statistics in Medicine, 38:2074–2102, 2019. doi: 10.1002/sim.8086.
3. A Burton, D G Altman, P Royston, and R L Holder. The design of simulation studies in medical statistics. Statistics in Medicine, 25:4279–4292, 2006. doi: 10.1002/sim.2673.
4. M Deforth, C E Gebhard, S Bengs, K P Buehler, R Schuepbach, A Zinkernagel, S D Brugger, C T Acevedo, D Patriki, B Wiggli, R Twerenbold, G Kuster, H Pargger, J Schefold, T Spinetti, P Wendel-Garcia, D Hofmaenner, B Gysi, M Siegemund, G Heinze, V Regitz-Zagrosek, C Gebhard, and U Held. Development and validation of a prognostic model for the early identification of COVID-19 patients at risk of developing common long COVID symptoms. Diagnostic and Prognostic Research, 6(1):22, 2022. doi: 10.1186/s41512-022-00135-9.

ID: 362 / **Poster 2: 5**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Biomarkers and diagnostics
Keywords: imaging biomarker, random variability, sample size calculation, sensitivity, specificity, Poster ID T5

A statistical framework for planning and analysing test-retest studies for repeatability of quantitative biomarker measurements

Moritz Fabian Danzer[1], **Maria Eveslage**[1], **Dennis Görlich**[1], **Benjamin Noto**[2]
[1]Institute of Biostatistics and Clinical Research, University of Münster, Germany; [2]Clinic for Radiology, University Hospital Münster, Germany; moritzfabian.danzer@ukmuenster.de

There is an increasing number of potential biomarkers that could allow for early assessment of treatment response or disease progression. However, measurements of quantitative biomarkers are subject to random variability. Hence, differences of a biomarker in longitudinal measurements do not necessarily represent real change but might be caused by this random measurement variability. Before utilizing a quantitative biomarker in longitudinal studies, it is therefore essential to assess the measurement repeatability. Measurement repeatability obtained from test-retest studies can be quantified by the repeatability coffiecient (RC), which is then used in the subsequent longitudinal study to determine if a measured difference represents real change or is within the range of expected random measurement variability. The quality of the point estimate of RC therefore directly governs the assessment quality of the longitudinal study.

RC estimation accuracy depends on the case number in the test-retest study, but despite its pivotal role, no comprehensive framework for sample size calculation of test-retest studies exists. To address this issue, we have established such a framework, which allows for flexible sample size calculation of test-retest studies, based upon newly introduced criteria concerning assessment quality in the longitudinal study. This also permits retrospective assessment of prior test-retest studies.

ID: 382 / **Poster 2: 6**
Poster Submission
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), High dimensional data, genetic and x-omics data, Real world data and evidence
Keywords: Spline Regression, UVA UVB Radiation, Skin cells, RNAseq-/Gene-Data, Poster ID T6

Interaction effects of UVA with UVB irradiation at the gene expression level in human skin cells

Yassine Talleb[1], Katharina Rolfes[2], Jochen Dobner[2], Andrea Rossi[2], Thomas Haarmann-Stemmann[2], Jean Krutmann[2], Katja Ickstadt[1]

[1]TU Dortmund, Germany; [2]IUF Düsseldorf, Germany; yassine.talleb@tu-dortmund.de

Ultraviolet (UV) B radiation (290-315 nm) and UVA (315 – 400 nm) are complete carcinogens and both are well known to contribute to the development of skin cancer in humans. Under physiological conditions, human skin is exposed to a mixture of UVB and UVA radiation from natural sunlight. Previous research, however, has primarily focused only on the effects of each type of radiation separately whereas the knowledge of the UVA-UVB-interaction under simultaneous exposure and its impact on human skin is quite unclear. In a previous photocarcinogenesis study we have found that simultaneous exposure of murine skin to UVB and non-carcinogenic doses of UVA increased UVB-induced photocarcinogenesis. These results indicate that detrimental effects caused by UVB and UVA radiation, if applied simultaneously, can enhance each other, even when the UVA dose per se does not cause significant skin damage. In the present study we would like to further analyze the interaction of UVB and UVA radiation, if applied simultaneously, by analyzing gene expression responses (by bulk RNAseq) in human skin cells. We are particularly interested in identifying the minimum UVA dose which is required to enhance UVB-induced skin damage.

The dose-response relationships involved will be examined more closely by using already existing RNAseq data sets of skin cells. Keratinocytes were irradiated either with only UVB or UVA, or with a combination (simultaneous irradiation) of both. The RNA of the cells was isolated at two different time points after irradiation (incubation time). Additional data was acquired for further incubation times and different physiologically relevant dosages and different ratios of UVA and UVB radiation. These settings were chosen to best support the statistical analysis.

When modelling univariate dose-response relationships separately for each incubation time, the information shared across the incubation times is not considered. However, it is critical to identify certain features of the relationship. We will consider data of all incubation times at once by modelling a two-dimensional surface with the help of tensor product B-splines. The identification of the minimum UVA dose, at which the combination irradiation shows a significantly different effect compared to UVB irradiation alone, will ultimately also be investigated and described with the help of spline regression models.

ID: 385 / **Poster 2: 7**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Biomarkers and diagnostics, Time-to-Event Analysis
Keywords: biomarker, surrogacy, time-to-event, Poster ID T7

Evaluate the association between longitudinal biomarker data and a time-to-event endpoint

Sauvageot Nicolas, Hsu Schmitz Shu-Fang
Statistics and Decision Sciences, Actelion Pharmaceuticals Ltd., Janssen Pharmaceutical Companies of Johnson & Johnson, Allschwil, Switzerland; NSauvage@its.jnj.com

Clinical trials often use a time-to-event endpoints to measure clinical outcome, with event status potentially changing with follow-up time. Biomarkers that are potentially associated with the clinical outcome are also often assessed longitudinally over time. Naturally, the association between the biomarker and the time-to-event endpoint is of special interest. Further interest is to evaluate whether the biomarker can serve as a surrogate endpoint.

In this context, the values of the biomarker observed at a specific timepoint are frequently used as a fixed covariate in a Cox proportional hazards model (Model 1) to analyse its impact on the time-to-event endpoint. Such a model ignores a large amount of biomarker information collected in the longitudinal process. A potentially more appropriate model is to incorporate the biomarker information as a time-varying covariate in the Cox model (Model 2) that implicitly assumes a step function for the biomarker trajectory due to lack of continuously monitored biomarker data. The biomarker value at a scheduled assessment timepoint is assumed constant up to the next assessment timepoint, i.e. the last-observation-carried-forward (LOCF) approach is applied. The resulting step function might not provide a good approximation of the true biomarker trajectory and could result in biased estimate towards zero (Arisido, 2019).

Joint modelling of the longitudinal biomarker data and time-to-event data with shared random effects could be a preferred approach (Model 3), which also allows for the inference on the association between the biomarkers longitudinal process and the hazard of the event. Previous studies showed that such models provide an unbiased estimate, showing improvement over the time-varying Cox model (Arisido, 2019). Moreover, such models can also be used to assess criteria laid out for a surrogate endpoint (Prentice 1989) and compute the proportion of treatment effect explained by a biomarker (Freedman 1992).

As an illustrative example, we applied different models to the data of a randomized double-blind, placebo-controlled phase 3 study. The biomarker of interest was measured at week 4, 8, 16, 26 and 52. The clinical outcome of interest was time to disease progression. Results were compared between models with respect to the association between the biomarker and the hazard of disease progression as well as the potential surrogacy of the biomarker for disease progression.

ID: 391 / **Poster 2: 9**
Poster Submission
Topic of Submission: Biomarkers and diagnostics
Keywords: sickle cell disease, biomarker, P-selectin, non-linear mixed effect model, Poster ID T9

Exposure - biomarker analysis in sickle cell disease patients

Kai Grosch
Novartis Pharma, Switzerland; kai.grosch@novartis.com

Crizanlizumab is a humanized monoclonal antibody against P-selectin for the prevention of vaso-occlusive crises in sickle cell disease (SCD). P-selectin (Psel)-mediated multi-cellular adhesion is a key factor in the pathogenesis of vaso-occlusion and vaso-occlusive crisis (VOC) as Psel, that is expressed on the surface of the endothelium, is thought to mediate abnormal rolling and static adhesion of sickle erythrocytes to the vessel surface. Shedding generates a soluble form of P-selectin that is absent of a transmembrane domain. We evaluated the relationship between crizanlizumab exposure from 5 mg/kg and 7.5 mg/kg dose and free, unbound soluble P-selectin levels, an exploratory biomarker, in sickle cell disease patients and described this relationship by a non-linear mixed effect model. This model supports dose and regimen decisions of crizanlizumab in sickle cell disease patients.

ID: 392 / **Poster 2: 10**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Free Contributions
Keywords: teaching, R, reproducibility, code, Poster ID T10

Teaching R to medical researchers

Monika Hebeisen, Stefanie von Felten, Ulrike Held
University of Zurich, Switzerland; monika.hebeisen@uzh.ch

There is an ongoing debate about deficiencies in the reporting of findings and reproducibility across different areas of medical research [1, 2]. One reason for non-reproducibility is the lack of statistical code to accompany the published research [3, 4]. At the Department of Biostatistics at the University of Zurich, the statistical programming language R is taught to medical students and senior medical researchers of the medical faculty since 2019.

In short introductory R courses of seven hours, participants learn to prepare data for analysis, to compute descriptive statistics, to perform simple statistical tests and to create graphics, using data examples published in the medical field. Upon completion of the course, participants can book code clinic sessions where project-specific questions are solved in one-to-one sessions.

We further teach dynamic reporting with R Markdown to medical students, in a semester course, again with applications of statistical methods in clinical research. Students learn to compile R Markdown reports under a predefined structure related to manuscripts in medical research, including tables and figures generated within R, references, and R package versions. The students appreciate the improved organization of material for their master or dissertation projects, because all important information is stored within one document.

Feedback of the participants is routinely recorded immediately at the end of the courses and in a follow-up survey that will be conducted in April 2023.

Between August 2020 and January 2023, 196 participants of the medical faculty completed the short introductory R course, of whom 174 (89%) filled in a feedback form right after the course. 64 (37%) graded the course as “very good” and 94 (54%) as “good”. Most of the participants (152, 87%) stated they would use R in their next research project, 11 (6%) said maybe they would do so. We will present the survey answers and show details on sustainability of the course, and whether R was used for the analysis in subsequent research by the participants.

For early-stage researchers, knowledge of R in combination with R Markdown is becoming increasingly important. Teaching R and R Markdown to medical researchers is boosting transparency, and therefore reproducibility. This is an important step for better credibility and validity in medical research.

References:

1. Niven, D.J., et al., Reproducibility of clinical research in critical care: a scoping review. BMC Med, 2018. 16(1): p. 26.
2. Wang, S.V., et al., Reproducibility of real-world evidence studies using clinical practice data to inform regulatory and coverage decisions. Nat Commun, 2022. 13(1): p. 5126.
3. DeBlanc, J., et al., Availability of Statistical Code From Studies Using Medicare Data in General Medical Journals. JAMA Intern Med, 2020. 180(6): p. 905-907.
4. Localio, A.R., et al., Statistical Code to Support the Scientific Story. Ann Intern Med, 2018. 168(11): p. 828-829.

ID: 395 / **Poster 2: 11**
Poster Submission
Topic of Submission: Estimands and causal inference
Keywords: Equivalence, Estimands, Poster ID T11

Interpreting ICHE9-R1 Guidance on Estimands in Equivalence Clinical Trials

Joëlle Monnet
Fresenius Kabi, Switzerland; joelle.monnet@fresenius-kabi.com

ICH adopted the ICH E9-R1 guidance in November 2019. Since the guideline was released for public consultation in 2017, its implementation, especially in superiority trials, has been discussed and described in the literature. However, the implementation of estimands in the context of equivalence trials is not covered in detail by the guideline itself, and less has been published in this context.

This poster will illustrate how the ICH E9-R1 estimand guideline can be interpreted and implemented in equivalence studies. We will present how the guideline was applied to the definition of estimands for a biosimilar equivalence study conducted to demonstrate the therapeutic equivalence between a proposed biosimilar and its corresponding reference product. As a general principle, the assumptions related to the mechanism of missingness, as well as the strategies proposed to deal with the pre-identified intercurrent events were selected with the objective to define sensitive estimands, allowing to detect treatment difference if any.

In particular, the “per-protocol analysis set” analyses were revisited to allow for the Estimand framework. A “Hypothetical continuing per Protocol Estimand” was defined, where data possibly impacted by an intercurrent event were censored and imputed as if the subjects would have continued to follow the protocol.

Definition, analysis, and some results of estimands constructed with an equivalence objective will be presented and discussed.

ID: 399 / **Poster 2: 12**
Poster Submission
Topic of Submission: High dimensional data, genetic and x-omics data
Keywords: genetic linkage analysis, microbiome, IBD, Poster ID T12

Network-based quantitative trait linkage analysis of microbiome composition in inflammatory bowel disease families

Arunabh Sharma, Olaf Junge, Silke Szymczak, Malte Rühlemann, Janna Enderle, Stefan Schreiber, Matthias Laudes, Andre Franke, Wolfgang Lieb, Michael Krawczak, Astrid Dempfle
University of Kiel, Germany; sharma@medinfo.uni-kiel.de

Introduction: Inflammatory bowel disease (IBD) is characterized by a dysbiosis of the gut microbiome that results from the interaction of the constituting taxa with one another, and with the host. At the same time, host genetic variation is associated with both IBD risk and microbiome composition.

Methods: In the present study, we defined quantitative traits (QTs) from modules identified in microbial co-occurrence networks to measure the inter-individual consistency of microbial abundance and subjected these QTs to a genome-wide quantitative trait locus (QTL) linkage analysis.

Results: Four microbial network modules were consistently identified in two cohorts of healthy individuals, but three of the corresponding QTs differed significantly between IBD patients and unaffected individuals. The QTL linkage analysis was performed in a sub-sample of the Kiel IBD family cohort (IBD-KC), an ongoing study of 256 German families comprising 455 IBD patients and 575 first- and second-degree, non-affected relatives. The analysis revealed five chromosomal regions linked to one of three microbial module QTs, namely on chromosomes 3 (spanning 10.79 cM) and 11 (6.69 cM) for the first module, chr9 (0.13 cM) and chr16 (1.20 cM) for the second module, and chr13 (19.98 cM) for the third module. None of these loci have been implicated in a microbial phenotype before.

Discussion: Our study illustrates the benefit of combining network and family-based linkage analysis to identify novel genetic drivers of microbiome composition in a specific disease context.

ID: 402 / **Poster 2: 13**
Poster Submission
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), Time-to-Event Analysis
Keywords: Poster ID T13

Comparison of different survival analysis models for estimation of time-to-hip fracture with death as a competing risk

Oluwabukunmi Mercy Akinloye[1], Sandra Freitag-Wolf[1], Astrid Dempfle[1], Claus-Christian Glüer[2]
[1]Institute of Medical Informatics and Statistics, Kiel University, Germany; [2]Department of Radiology and Neuroradiology, Kiel University, Germany; stu232964@mail.uni-kiel.de

Prediction of future fractures and identification of individuals with a higher risk of refracture is essential for the prevention and treatment of osteoporosis [1]. With survival analysis, the time to a fracture event can be assessed. However, a distinction needs to be made regarding study participants for whom the event of interest is not observed either due to classic censoring or due to a competing event such as death. We, therefore, compared three survival analysis models, the Cox Proportional Hazard (PH), the Cause-Specific Cox (CSC), and the Fine-Gray, for estimating time to osteoporotic hip fracture while considering death as a competing risk.

Using a subset of the Study of Osteoporotic (SOF) dataset, we focused on three key variables: age, BMI, and previous fracture (PFX). These three variables were then used to fit two Cox PH models for the events of incident hip fracture and death. Older age and history of PFX, are associated with a decreased time to incident hip fracture. In contrast, a larger BMI seems to confer a slightly protective effect. In the Cox PH model for the event of death, all three variables are positively related to earlier deaths. When comparing the output of the Cox models to the CSC model, we see very minor differences between the hazard ratios calculated from both methods. However, with FG for incident hip fracture, we see a decrease in the hazard ratios for age and previous fracture and a slight increase in the hazard ratio for BMI.

While our results indicate some differences between the models, these differences are mostly minor. Currently, we are extending this study to include more variables, the full SOF dataset, and the use of other machine learning methods with the goal of refining and utilizing these models for clinically applicable risk prediction.

Reference
Morin, S. N., Lix, L. M., & Leslie, W. D. (2014). The importance of previous fracture sites on osteoporosis diagnosis and incident fractures in women. Journal of Bone and Mineral Research, 29(7), 1675-1680.

ID: 417 / **Poster 2: 14**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), Prevention and handling of missing data, Machine Learning and Data Science
Keywords: missing data, multiple imputation, predictive modelling, Poster ID T14

Efficiently involving clinical experts for handling missing data: a case study

Rieke Alpers[1,2,3], Sebastian Daniel Boie[4], Eduardo Salgado[4,5], Felix Balzer[4], Pamela Bendz[6], Sophia Schmee[6], Anja Hennemuth[1,2,3], Markus Hüllebrand[1,2,3], Max Westphal[1]

[1]Fraunhofer Institute for Digital Medicine MEVIS; [2]Deutsches Herzzentrum der Charité, Institute of Computer-assisted Cardiovascular Medicine; [3]Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin; [4]Institute of Medical Informatics, Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health; [5]Department of Anesthesiology and Intensive Care Medicine, Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health; [6]Department of Anaesthesiology, Intensive Care, Emergency and Pain Medicine, University Hospital Wuerzburg; rieke.alpers@mevis.fraunhofer.de

Encountering missing values is almost inevitable when working with clinical data. Among the variety of methods available for dealing with them, experts often favour multiple imputation over simpler techniques [1]. However, recent literature reviews reveal that clinical risk prediction models show poor handling of missing data and reporting thereof: Most authors either delete missing values or deploy mean imputation, or they rather omit reporting on missingness in their data completely [2]. Why is the problem of missing data widely discussed in theory but seldom tackled in practice? Reasons might be that complex imputation models like MICE are more error-prone and time-consuming and may require consultation with clinical experts. This work aims to answer whether the additional effort of including medical knowledge for imputation results in improved prediction models.

The study uses routine data from ten thousand patients undergoing elective surgery from 2016 to 2022 at the university hospital Charité in Berlin. The dataset includes demographics, vital signs, clinical scores, laboratory values, comorbidities, and surgery information as well as indicators for eight common perioperative complications (e.g., pneumonia, bleeding, or death). For enabling a comparison of missing data handling methods, we developed prediction models for the eight complications by training the methods crosswise with different Machine Learning algorithms. Standard techniques for deletion of missing values, simple imputation, and multiple imputation were compared amongst each other and to new procedures in which each feature receives its own imputation model based on clinician's appraisals. For this purpose, we performed interviews with clinicians from two different sites. In the analysis, we assessed how much time was needed to specify and train the imputation models. We measured it against the imputation's effects on prediction performance and the clinician's trust in the predictions, knowing how missingness is handled in new data.

In our cohort, missing data occurred in all patients and a quarter of the features, with an overall missing ratio of around 15%. The prediction models performed similarly for all missing data handling methods. At the cost of a longer time to specify the imputation models, the procedures involving clinical experts increased trust in the final predictions compared to other methods. We will validate our findings further and assess the effect of altering the imputation methodology on regional and temporal generalizability with data from an ongoing prospective study including four hospitals. We expect that standard methods for handling missing data may overemphasize cohort-specific characteristics of the data, whereas embedding expert knowledge may help to better translate to the broader patient population.

References

1. Jakobsen, J.C., Gluud, C., Wetterslev, J. et al. (2017). When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. *BMC Med Res Methodol,* 17, 162. https://doi.org/10.1186/s12874-017-0442-1
2. Nijman, S.W.J., Leeuwenberg, A.M., Beekers, I. et al. (2022). Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review. *JCE*, 142, 218-229. https://doi.org/10.1016/j.jclinepi.2021.11.023.

ID: 420 / **Poster 2: 15**
Poster Submission
Topic of Submission: Epidemiology
Keywords: Short-term air pollution exposure, particulate matter, transepidermal water loss, Bayesian hierarchical model, Poster ID T15

Short-term effects of particulate matter on transepidermal water loss in elderly Caucasian women

Claudia Wigmann, Tamara Schikowski
IUF - Leibniz Research Institute for Environmental Medicine, Germany; claudia.wigmann@iuf-duesseldorf.de

Transepidermal water loss (TEWL) describes the loss of water through the epidermis via diffusion or evaporation and is important for the evaluation of the skin barrier function. TEWL is known be altered, for example, in physically damaged or eczema-affected skin. In addition, it was shown to be influenced by age as well as environmental factors such as temperature and humidity. Recent studies also suggested short-term effects of air pollutants on TEWL.

In the longitudinal Study on the influence of Air pollution on Lung function, Inflammation and Aging (SALIA) we investigated the effects of short-term exposure to particulate matter with a diameter of 2.5 micrometers or less ($PM_{2.5}$) on TEWL. During a follow-up investigation in 2018/2019 including 224 women aged 75-89 years the TEWL was measured in eight locations across the body. Exposure to $PM_{2.5}$ at the participant's home address on the day of the investigation was assessed via chemical transport models of the German Environment Agency UBA.

Since levels of TEWL differ between different body parts we used a hierarchical Bayesian regression model, which allows for possibly different effects of $PM_{2.5}$ among the eight locations while at the same time estimating an overall effect of $PM_{2.5}$ on TEWL. The model includes random participant effects to account for the repeated measurements and it was adjusted for ambient temperature, humidity and personal confounder variables. Estimates were derived as percentage change with respect to an IQR increase in $PM_{2.5}$.

We found higher levels of TEWL on the forehead (posterior median, 95% credible interval: 1.49, [1.25, 1.90]) and on the upper side of the hand (1.22, [1.02, 1.54]) compared to the overall mean. In addition, TEWL was increased after short-term exposure to $PM_{2.5}$ on the wrist (1.25 [1.12, 1.39]) and the cheek (1.11, [1.002, 1.22]). The overall effect of $PM_{2.5}$ on TEWL was estimated as 1.07 [0.98, 1.19].

ID: 422 / **Poster 2: 16**
Poster Submission
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), Epidemiology
Keywords: reference interval, standard deviation score, outliers, Poster ID T16

Correction for outlier removal in parametric estimation of reference intervals and standard deviation scores

Andreas Gleiss
Medical University of Vienna, Austria; andreas.gleiss@meduniwien.ac.at

Reference intervals and standard deviation scores ('z scores') are widely used as diagnostic tools in various biomedical fields. They are applied to laboratory parameters in clinical chemistry, psychometric tests in neurology, or parameters of children's growth in pediatrics. Usually, samples from a 'normal' population form the data basis for the estimation of reference distributions. This target population may or may not include unhealthy individuals depending on whether the reference should reflect the healthy or the general population. Health status, however, is not collected in some types of reference studies, e.g. in pediatrics.

Parametric estimation of references distributions is complicated by extreme values that may be outliers relative to the assumed (working) distribution or model, even for the unconditional case (i.e. without dependence on covariables such as age). If sample size is moderate, genuine outliers may by chance be over-represented in the sample used for estimating the reference distribution and impair the selection of a suitable model. Second, if the target population is to include unhealthy individuals with their representative share, the reference distribution to be estimated is a mixture of a major 'healthy' part plus a 'contamination'. Finally, the sample may be contaminated by observations that are not members of the target population but remain undetected.

Often in practice, the origin of the extreme values at hand is unknown, but the interest is in one or both of the extreme tails of the distribution. However, most ad-hoc methods proposed in the literature for outlier removal or correction are only designed for one of the cases. In this contribution, the simple approaches of ignoring, deleting or winsorizing outliers are compared to two ex-post correction methods, one of which is newly proposed. Correction methods remove outliers before estimating the reference distribution, thus allowing to select a less complex model; for new observations from individuals to be diagnosed, percentiles are calculated based on the estimated reference distribution and then corrected for the removal of outliers by appropriate rescaling and shifting.

Simulation studies with normal, skewed and heavy-tailed distributions, varying sample size and varying degree of contamination show the strengths and limitations of the considered methods. The simple ad-hoc methods highly depend on their respective assumptions. While the first correction method over-corrects the deletion of outliers, the new correction method adequately corrects for genuine outliers being removed and attenuates the effect of contamination. Data on body mass index and body proportions from a large Austrian pediatric study are used for demonstration.

ID: 426 / **Poster 2: 17**
Poster Submission
Topic of Submission: Biomarkers and diagnostics, Statistical modelling (regression modelling, prediction models, …)
Keywords: diagnostic studies, clustered data, factorial designs, covariate adjustment, Poster ID T17

Covariate adjustment, factorial designs and clustered data in diagnostic studies

Philipp Weber[1], **Katharina Kramer**[2], **Antonia Zapf**[1]
[1]University Medical Center Hamburg-Eppendorf, Germany; [2]University of Augsburg, Germany; p.weber@uke.de

Diagnostic tests are commonly evaluated by estimating the area under the receiver operating characteristic curve (AUC), as well as sensitivity and specificity at given diagnostic cut-offs. One difficulty with diagnostic trials is that many of them use factorial designs. This means that different combinations of readers and methods may be used to diagnose a patient. In addition, diagnostic studies may generate clustered data by repeated measurements over time or several lesions. See [1, Lange] for a mathematical framework to deal with both of these difficulties.

Additionally, it may be of interest to correct the estimation procedure of the above mentioned accuracy measures for covariates. For example, it may be the case that age, weight or height influence the diagnostic accuracy of a test. In [2, Zapf] a methodological approach is presented to adjust the AUC for such covariates, while also allowing for factorial designs. We developed a modification of the adjustment approach to guarantee unbiased estimators also for sensitivity and specificity.

In this talk, we present the new approach and give a short overview of the corresponding R package (under development).

References:

1. Lange, K. (2011, March 4). *Nichtparametrische analyse diagnostischer Gütemaße bei Clusterdaten*. Retrieved February 27, 2023, from http://dx.doi.org/10.53846/goediss-3538
2. Zapf, A. (2009, October 23). *Multivariates nichtparametrisches Behrens-Fisher-problem MIT Kovariablen*. Retrieved February 27, 2023, from http://dx.doi.org/10.53846/goediss-2488
3. Lange, K., & Brunner, E. (2012). Sensitivity, specificity and ROC-curves in multiple reader diagnostic trials—a unified, nonparametric approach. *Statistical Methodology*, *9*(4), 490–500. https://doi.org/10.1016/j.stamet.2011.12.002

ID: 430 / **Poster 2: 18**
Poster Submission
Topic of Submission: Estimands and causal inference, Epidemiology, Real world data and evidence
Keywords: directed acyclic graphs, Poster ID T18

Directed acyclic graph-based data simulations in R/dagR: current state and open issues

Lutz Philipp Breitling
University of Heidelberg, Germany; l.breitling@posteo.de

Directed acyclic graphs (DAG) have become a well established tool to investigate confounding and bias in observational studies and real-world data. Simulating data according to the causal structures of a given DAG could potentially be useful for teaching, methods research, deciding on appropriate analytical approaches, and even during study design. Convenient simulation functionalities nonetheless remain limited in published software.

The R package dagR, which can also be used to identify (minimal) sufficient adjustment sets, initially included the possibility to simulate data for a given DAG including any combination of binary and continuous nodes/variables, with dependencies conforming to a logistic or linear model, respectively [1,2]. Functionalities to simulate binary variables based on a risk difference specification were added more recently [3].

Some examples are presented to demonstrate potential applications of the package in teaching and research.

Work in progress includes the implementation of time-to-event and repeated measurements simulations, which could pave the way to much wider application in fields of current interest, such as emulated target trial methodology.

References:

1. Breitling LP, Duan C, Dragomir AD, Luta G. Using dagR to identify minimally sufficient adjustment sets and to simulate data based on directed acyclic graphs. Int J Epidemiol 2021;50(6):1772–1777
2. Duan C, Luta G, Dragomir AD, Breitling LP. Reflection on modern methods: understanding bias and data analytical strategies through DAG-based data simulations. Int J Epidemiol 2021;50(6):2091–2097
3. dagR 1.2.1 (2022-10-09), https://cran.r-project.org/package=dagR

ID: 432 / **Poster 2: 19**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), Time-to-Event Analysis, Statistical hypothesis testing (covariate adjustment, nonparametric methods, multiple comparisons, …)
Keywords: Discrete time simulation, Multiple Sclerosis, Interaction, Poster ID T19

Modeling the interaction of two time-dependent covariates – a simulation study based on pregnancy data in Multiple Sclerosis

Marianne Charlotte Tokic[1], Sandra Thiel[2], Kerstin Hellwig[2], Nina Timmesfeld[1]
[1]Department of Medical Informatics, Biometry and Epidemiology, Ruhr University Bochum, Germany; [2]Department of Neurology, St. Josef-Hospital-Katholisches Klinikum Bochum, Ruhr University Bochum, Germany; tokic@amib.rub.de

In the context of multiple sclerosis (MS), relapse rates are seen as a key indicator of disease activity. Due to its gynecotropia and the peak of initial manifestation in child-bearing years[1], MS-disease management during pregnancy is of critical importance. A study by Confavreux et al.[2] found that pregnancy reduced relapse rates in mainly untreated MS patients. However, the use of modern disease-modifying treatments (DMTs) adds complexity to pregnancy planning. Most modern biologics are not licensed for use during pregnancy and must be discontinued prior to conception or in early pregnancy.

However, upon discontinuation of DMTs, the issue of rebound relapses arises. Natalizumab, for example, has been found to increase relapse rates in 3 months post-cessation exceeding pre-medication levels [3]. It is yet to be determined, whether the protective effect of pregnancy also applies to rebound relapses. Current observational studies remain largely inconclusive[4,5], partly due to the complexity of the temporal relationship between pregnancy duration and the cessation of the disease-modifying therapy (DMT) relative to conception.

To further explore the clinical question of whether pregnancy reduces the intensity of rebound relapses, we must consider the interaction between these two time-dependent covariates. Therefore, in this study we use simulated data to examine which analysis methods are best suited to capture this pregnancy*rebound interaction, in terms of power, bias, and communicability.

We conducted a discrete-time simulation[6] [DTS], informed by clinical data, to generate 1000 repeat samples of differing parameter sets with varying sample sizes of patients treated with a generic rebound-inducing DMT. We modeled the time-varying effects of the DMT and pregnancy as discrete and continuous effects.

To test the interaction hypothesis, we used forms of recurrent event Cox regression and Poisson regression models. The models are compared based on bias and power of the interaction term only. We will present results and discuss the implications for further studies in this field.

References:

1. Holstiege J, Steffen A, Goffrier B, Bätzing J. Epidemiologie der Multiplen Sklerose – Eine populationsbasierte deutschlandweite Studie. 2017 Dec 7;
2. Confavreux C, Hutchinson M, Hours MM, Cortinovis-Tourniaire P, Moreau T. Rate of Pregnancy-Related Relapse in Multiple Sclerosis. N Engl J Med. 1998 Jul 30;339(5):285–91.
3. Papeix C, Vukusic S, Casey R, Debard N, Stankoff B, Mrejen S, et al. Risk of relapse after natalizumab withdrawal: Results from the French TYSEDMUS cohort. Neurol Neuroimmunol Neuroinflamm. 2016 Dec;3(6):e297.
4. Hellwig K, Tokic M, Thiel S, Esters N, Spicher C, Timmesfeld N, et al. Multiple Sclerosis Disease Activity and Disability Following Discontinuation of Natalizumab for Pregnancy. JAMA Network Open. 2022 Jan 24;5(1):e2144750.
5. Portaccio E, Moiola L, Martinelli V, Annovazzi P, Ghezzi A, Zaffaroni M, et al. Pregnancy decision-making in women with multiple sclerosis treated with natalizumab: II: Maternal risks. Neurology. 2018 Mar 6;90(10):e832–9.
6. Tang J, Leu G, Abbass HA. Discrete Time Simulation. In Simulation and Computational Red Teaming for Problem Solving (eds J. Tang, G. Leu and H.A. Abbass). 1st ed. Wiley; 2019 . https://doi.org/10.1002/9781119527183.ch8

ID: 438 / **Poster 2: 20**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Estimands and causal inference, Epidemiology
Keywords: collider stratification bias; obesity paradox; simulation, Poster ID T20

Collider stratification bias as a potential explanation for the obesity paradox – a simulation analysis

Josef Fritz[1,2], **Tanja Stocks**[1]
[1]Lund University, Sweden; [2]Medical University of Innsbruck, Austria; josef.fritz@med.lu.se

Collider stratification bias is routinely brought up as a potential explanation when the association of obesity with the development of a disease is of different magnitude than the association of obesity with disease outcome. The extreme case, where obesity is a risk factor for the disease, but is associated with improved survival in the diseased, as observed for example for cardiovascular disease, diabetes, chronic kidney disease, and several types of cancer, is known as the “obesity paradox”. Prerequisites for collider bias to occur are an effect of obesity on the development of disease, and an unadjusted confounder of the disease development and outcome relationship, such as for example (unknown) genetic risk factors. Simulating multiple scenarios with different choices of input parameters (*i.e.* effect of obesity on disease development; strength of the unadjusted disease development and outcome confounder; disease and outcome incidence), we investigated under which scenarios the danger of collider bias is real, and under which it is negligible. A main finding was that the potential impact of collider bias is relatively unaffected by the specific prevalences, but sensitive on the magnitude of the effect of obesity on disease development. As long as this effect is not too large (hazard ratio (HR) between 0.67 and 1.5), even assuming a strong background confounder, collider bias is relatively small, meaning that the true, unbiased HR is distorted by less than 5% on the HR scale (*i.e.* less than 0.05 on the absolute HR scale if the HR is not too big). For comparison, in large observational studies with a couple of hundred thousand participants, 95% confidence intervals for mortality outcomes for a rare, but not too deadly, disease often cover a range of 0.2 or more on the HR scale, and thus the random variability in HR estimates outweighs the likely magnitude of collider bias. Only in case of small causal effects on the disease outcome, collider bias is able to fully explain the obesity paradox. Classical confounding, detection bias, heterogeneity of disease bias, or model misspecification bear a much higher potential of substantially distorting observed associations than collider stratification in many typical scenarios.

ID: 441 / **Poster 2: 21**
Poster Submission
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), Machine Learning and Data Science, High dimensional data, genetic and x-omics data
Keywords: compositional data analysis, high dimensional data, Poster ID T21

Interpretable two-stage predictive modeling of compositional microbiome data

Minh Viet Tran[1,2,3], **Christian L. Müller**[1,2,3,4]
[1]Ludwig-Maximilians-Universität München, Germany; [2]Helmholtz Munich, Germany; [3]Munich Center for Machine Learning, Germany; [4]Center for Computational Mathematics, Flatiron Institute, USA; viet.tran@campus.lmu.de

High-throughput amplicon and metagenomics sequencing data have become an invaluable resource for assessing broad statistical patterns of associations between microbial communities and their environment. Host information and a priori expert knowledge represented on a tree are often provided along with the high-dimensional and compositional microbial data. Sparse and compositional aware predictive methods incorporating the hierarchical information and other (non-compositional and non-hierarchical) host-related features are therefore desirable.

Here, we propose a two-stage classification framework encompassing and extending prior hierarchical expert knowledge to associate compositional microbial data with the host's trait, allowing for non-compositional covariates. The first stage extends the sparse log-contrast model to classification cases imposing a zero-sum constraint for compositional awareness and an l1 penalty for sparsity to the convex loss. Another ingredient is the tree aggregation idea allowing for flexible aggregation and selection along the hierarchical tree. The second stage simplifies the model by running another sparse method on log-ratio pairs transformed aggregated variables defined by the support of the first stage.

The applications used the taxonomic tree annotation of the microbes. The method was benchmarked against state-of-the-art black-box models on various datasets, revealing comparable predictive performance while being sparser. The analysis of two real-world datasets revealed the more fine-grained log-ratio Firmicutes (Phylum level) / Bacteroidales (Order level) as possible microbial biomarkers for irritable bowel syndrome instead of the often used fixed-level ratio Firmicutes / Bacteroidetes on the Phylum level.

The scalable and interpretable modeling framework for high-dimensional compositional microbiome data allows for selecting log-ratios pairs predictive of a host's trait. The peculiarity of the log-ratios is to allow for aggregation but to restrict those to a prior defined hierarchical structure. This framework contributes to the growing toolbox for high-dimensional compositional data analysis for microbial data.

ID: 444 / Poster 2: 22
Poster Submission
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), Time-to-Event Analysis, Statistical hypothesis testing (covariate adjustment, nonparametric methods, multiple comparisons, …)
Keywords: relational event models, time-varying effects, random effects, goodness of fit, alien species invasions, Poster ID T22

Evaluating the Goodness of Fit of Relational Event Models via Stratified Sums of Martingale Residuals

Martina Boschi, Ernst-Jan Camiel Wit
Università della Svizzera italiana, Switzerland; martina.boschi@usi.ch, ernst.jan.camiel.wit@usi.ch

Temporally ordered interactions between actors constitute a complex temporal network where past configurations may be partly responsible for future ones. These interactions are sometimes referred to as relational events. The relational event model (REM) seeks to capture the underlying dynamics driving these interactions. Nevertheless, assessing the goodness of fit (GOF) of these models is still mostly an open field of research, particularly for REMs with time-varying and random effects.

Our proposal relies on a cumulative martingale-residual process evaluated for a smooth mixed-effect REM estimated using case-control sampling. We may derive various GOF statistics, such as the normalized sum of martingale residuals stratified by receivers. Under the null hypothesis of the adequacy of the model, the recipient-specific test statistic can be shown to be asymptotically standard normally distributed.

In an empirical application, we fit a smooth case-control REM to sequences of alien species invasions. A first record, defined as the first year in which a species is detected as alien in a region where it was not native, is our relational event of interest. To test the GOF of our proposed model, we compute the country-specific sum of martingale residuals. With a significance level of 5%, there is no evidence of model misspecification related to the regions.

Martingale residual stratification may be done in a variety of ways. Our approach can easily be extended to determine whether any other network dynamics features have been adequately incorporated into the model.

ID: 321 / **Poster 2: 23**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Free Contributions
Keywords: Accessibility, disability, inclusive design, figures, tables, Poster ID M22

Statistics - made accessible

Ursula Becker
F. Hoffmann-La Roche, Switzerland; ursula.becker@roche.com

As statisticians we have been making many efforts to render our work more accessible for non-statisticians. One aspect however which has been neglected so far is the topic of Digital Accessibility. What can we statisticians do in order to make our work also accessible for people with permanent or temporary disabilities? How can we change the way we present data in order to be inclusive?

Digital Accessibility (DA) is an integral element in creating an inclusive (digital) world and is closely linked to the topic of the conference “From Data to Knowledge. Advancing Life Sciences.”: DA ensures information can be reached, used and understood by everyone..

An estimated 1.3 billion people – or 1 in 6 people worldwide – experience significant disability (WHO, 2022). DA is crucial so that these people but also many of us who do not consider ourselves being disabled can access digital content. Disabilities can be visible, but many disabilities are invisible. The most common categories are

- Visual disabilities (e.g., blindness, low vision, color blindness)
- Auditory (e.g., deaf and hard of hearing)
- Motor (e.g., inability to use a mouse, slow response time, limited motor control)
- Cognitive (e.g., inability in memory, attention/focus, text processing)

In order to make content accessible, it should be easy to see, easy to hear, easy to interact with and easy to understand.

Ultimately, DA means good design and good usability and many other people without permanent disabilities will also benefit: people not fluent in a language, older people, people with “temporary disabilities” due to accident or illness, people with “situational limitations” such as weather-based conditions (bright sunlight, loud environment etc.) as well as people using a slow internet connection or legacy browsers or unable

In our working context DA is very relevant in interactions with our stakeholders (patients, physicians, researchers and students, existing and future employees, investors). It applies to almost everything we do: from choosing fonts and colors to creating tables and graphs which can be used for reports, publications, presentations, websites, social media etc.

In the session I would like to discuss concrete examples what can we do differently as statisticians in order to make our work digitally accessible, e.g.,

- Provide sufficient contrast
- Conscious use of colors (do not use colors alone to convey information)
- Ensure that interactive elements are easy to identify
- Provide clear and consistent navigation options
- Provide easily identifiable feedback
- Use headings and spacing to group related content

People should leave with tangible ideas on easy things to change in order to make our work more accessible and thus easier to understand and to spread.

Sources:

WHO 2022. Fact Sheet Disability. Available: https://www.who.int/news-room/fact-sheets/detail/disability-and-health [2023, February 28]

Abstracts for Oral Contributions

ID: 488 / STRATOS 2: 3
Presentation Submissions - Featured Session
Featured Sessions: Ten years of the STRengthening Analytical Thinking for Observational Studies (STRATOS) initiative – progress and looking to the future
Keywords: Simulations; Quantitative bias analysis; Survival analysis, Sensitivity analyses, Non- collapsibility, Time-varying exposure, Interval censoring

Data-Driven Simulations to Assess the Impact of Study Imperfections In Real-World Time-to-Event Analyses

Michal Abrahamowicz[1,5], Marie-Eve Beauchamp[1], Anne-Laure Boulesteix[2], Tim P. Morris[3], Willi Sauerbrei[4], Jay S. Kaufman[1]
[1]McGill University, Montreal, Canada, Canada; [2]Ludwig-Maximilians-Universität München, München, Germany; [3]University College London, London, UK; [4]Medical Center - University of Freiburg, Freiburg, Germany; [5]on behalf of the STRATOS Simulation Panel; michal.abrahamowicz@mcgill.ca

Quantitative bias analysis (QBA) permits assessment of the expected impact of various imperfections of the available data on the results and conclusions of a particular real-world study. This article extends QBA methodology to multivariable time-to-event analyses with right-censored endpoints, possibly including time-varying exposures or covariates. The proposed approach employs data-driven simulations, which preserve important features of the data at hand while offering flexibility in controlling the parameters and assumptions that may affect the results. First, the steps required to perform data-driven simulations are described, and then two examples of real-world time-to-event analyses illustrate their implementation and the insights they may offer. The first example focuses on the omission of an important time-invariant predictor of the outcome in a prognostic study of cancer mortality, and permits separating the expected impact of confounding bias from non-collapsibility. The second example assesses how imprecise timing of an interval-censored event – ascertained only at sparse times of clinic visits – affects its estimated association with a time-varying drug exposure. The simulation results also provide a basis for comparing the performance of two alternative strategies for imputing the unknown event times in this setting.

Thursday, 07/Sept/2023 11:40am - 12:00pm
ID: 253 / **S66: 4**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Advanced survival analysis
Keywords: antibody waning, frailty approach, Markov Chain Monte Carlo simulation, mixture models

Individual heterogeneity in humoral immune response: A Bayesian frailty approach

Steven Abrams[1,2], Adelino Martins[3], Niel Hens[1,2]
[1]UHasselt, Belgium; [2]University of Antwerp, Belgium; [3]Eduardo Mondlane University, Mozambique;
steven.abrams@uhasselt.be

Background

The analysis of multivariate serological data, i.e., Type I interval-censored or current status data coming from blood serum samples tested for the presence of antibodies against multiple pathogens, gained attention in recent years. More specifically, refinements towards accounting for individual heterogeneity in the acquisition of infections, non-immunizing infection dynamics and more complicated association structures have been proposed and studied in detail. Despite the common use of a so-called threshold approach to classify individuals as seronegative or -positive depending on a continuous antibody titer measurement for each pathogen under study, the subjective choice of a single or even multiple thresholds is a strong limitation. In this work, we focus on directly using individual-level bivariate continuous antibody titer data to estimate parameters related to the occurrence of two pathogens.

Methods

We consider a Bayesian bivariate mixture approach to model continuous antibody titer data on two pathogens in the presence of individual heterogeneity and to study the implied association in the acquisition of these two pathogens. Our approach extends a common frailty approach for paired current status data, which is subject to potential misspecification of thresholds as previously mentioned.

Data application

We fitted the aforementioned model to bivariate serological data on varicella-zoster virus (VZV) and parvovirus B19 (PVB19). Given recent evidence of possible reinfections with PVB19, we investigated processes of waning of humoral immunity by allowing for an age-dependent change in mean antibody titer concentration. We discuss these humoral immunity processes in more detail and consider different model choices in view of such processes.

Results and conclusions

The estimated seroprevalence for PVB19 is characterized by a steep increase with increasing age, following infections among young children, followed by a decrease between the age of 20 to 40 years after which the seroprevalence increases again. Moreover, the evolution of the mean antibody titer concentrations is rather constant across age groups, indicating that despite a decay in humoral immunity at the individual-level, population-level mean antibody titer values remain unchanged because of reinfections with PVB19 among 20-40 years old. Given the risk of spontaneous abortion after PVB19 infection during pregnancy, waning of humoral immunity in 20-40 years could be responsible for an excess of miscarriage and fetal death. For VZV, the seroprevalence is monotonically increasing, indicating that varicella infection is responsible for high levels of humoral immunity persisting for life. The mean antibody levels show a slight decrease with increasing age among seropositive individuals, however, not to an extent that seroprotection is not ensured for life. In general, based on our analyses, we showed that the mixture model provides additional insights concerning waning of IgG antibodies as compared to more traditional frailty approaches while the model is sufficiently flexible to capture observed dynamics in IgG antibodies. Furthermore, the model accounts for association in the acquisition of the pathogens under study through the specification of random effects termed frailties to explicitly link our approach to survival models that have been used in the past.

Monday, 04/Sept/2023 3:00pm - 3:20pm
ID: 356 / **S9: 3**
Presentation Submissions - Invited Session
Invited Sessions: A causal inference perspective on estimands in clinical trials
Keywords: Causal inference, estimands, target trial framework, principal stratum

What is the role of causal thinking in global drug development?

Mouna Akacha
Novartis Pharma AG, Switzerland; mouna.akacha@novartis.com

Causal thinking and related inference methods are gaining increasing prominence in global drug development in light of the recently published ICH E9(R1) guideline on estimands and sensitivity analysis (2019) and the FDA draft guideline on covariate adjustment (2021). These guidelines refer to terminology, concepts and methods from the causal inference literature, such as potential outcomes, principal stratification, non-collapsibility and standardization.

In this talk we build upon these recent developments by examining how causal inference provides a convenient mathematical language and tools to formally establish causal relationships between, e.g., drug and effect. We believe that causal inference methods can be used in many drug development settings, including those outlined in the two guidelines above, but also for the use of external control data and understanding cause and effect in pharmacometric and pharmacovigilance applications. We illustrate the importance and potential impact of causal inference by presenting two case studies. In the first case study, we will discuss how trial external data can be leveraged using the target trial and estimand frameworks for an oncology trial. In addition, the use of a principal stratum estimand to answer a clinical meaningful question in multiple sclerosis will be discussed.

Thursday, 07/Sept/2023 11:40am - 12:00pm
ID: 179 / **S68: 4**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Design of preclinical experiments
Keywords: animal studies, Bayes, factorial designs

Designing preclinical experiments using factorial and Bayesian approaches

Andreas Allgöwer, Benjamin Mayer, Theresa Unseld
Institute of Epidemiology and Medical Biometry, Ulm University, Germany; andreas.allgoewer@uni-ulm.de

Background:

The planning and analysis of preclinical animal trials is challenging because of small sample sizes and sparse prior data. Standard analysis results (e.g. from frequentist t-tests) are possibly of limited validity in calculating an appropriate sample size for the actual animal experiment. So alternative approaches are desired such as (i) a rearrangement of prior data to a final full factorial design or (ii) the incorporation of prior knowledge in Bayesian analysis to get a meaningful number of necessary animals.

Methods:

For the first approach, based on historical data of two groups, possible full factorial designs were created. Therefore different interaction patterns were presumed and sample sizes for the potential main and interaction effects were calculated. For the second approach, different concepts for design analysis with Bayesian methods were compared and evaluated with respect to their applicability in preclinical, translational research.

Results:

If a full factorial design is planned, the interaction effects should always be taken into account. Besides a gain in information, also a smaller sample size is possible. Bayesian methods allow a better representation of uncertainties in the model parameters. Using design analysis as basis for a sample size decision, instead of the classical sample size calculation by solving power equations may be preferred since it reflects the alignment to additional design goals and the sensitivity of the estimation results to the model assumptions.

Discussion:

As with a small group size, the usage of parametric test assumptions may not be valid. Moreover, current limitations reside in the complexity of Bayesian methodology, which make it challenging to understand and control the impact of single design components on estimation results, and in the limited availability of historical animal experiment data.

Thursday, 07/Sept/2023 12:00pm - 12:20pm
ID: 375 / S64: 5
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), Estimands and causal inference, Machine Learning and Data Science
Keywords: machine learning, internal-external validation, generalizability, reproducibility, transferability

An estimand framework to guide model and algorithm evaluation in predictive modelling

Rieke Alpers, Max Westphal
Fraunhofer Institute for Digital Medicine MEVIS; max.westphal@mevis.fraunhofer.de

The goal of evaluation studies in supervised machine learning (ML) is the quantification of the generalization capability of a trained predictive model or a learning algorithm. For an unbiased estimation, data splitting is mandatory as the apparent training performance is not indicative of the targeted out-of-sample performance [1]. Various designs and methods have been proposed for a valid and efficient estimation, e.g. different cross-validation variations (k-fold, leave-one-out, grouped, stratified, nested). For ML practitioners, it is however often unclear which of the available approaches is most suitable for their specific problem.

In this work, we connect this issue to the ongoing estimand discussion in statistics [2]. We performed a selective literature review to summarize common solutions and pitfalls in the context of clinical risk prediction modelling. We derived a new framework that can guide ML practitioners through their estimand definition. We also conducted a range of numerical experiments with real and simulated data to investigate the consequences of inconsistencies between the target estimand and the actual experimental design. Moreover, we developed a R new package to allow the direct implementation of our framework in practice.

Our derived estimand framework requires the characterization of the theoretical estimand ("What is the estimation target?") and the empirical estimand ("What is actually being estimated?") [3]. For this purpose, the relevant patient population(s) and differences between development and implementation context are described by a set of constraints (e.g. "model implemented in the same clinic(s) where the development data has been sampled" vs. "model implemented in new clinic(s)"). Deviations between the theoretical and empirical estimand, which cannot be avoided in practice, need to be rectified by (transferability) assumptions. This enables a precise description of the relevant estimand(s) and the limitations of the evaluation study (unrealistic transferability assumptions). The framework allows to improve the current practice in the clinical risk prediction literature where the estimand is often only vaguely defined and the validity and usefulness of the performance estimate(s) is thus unclear. Our numerical experiments indicate that inconsistencies between theoretical and empirical estimand can lead to a severely biased performance estimation. To mitigate this issue in future ML evaluation studies, our new R package 'mldesign' provides a simple interface to transfer an estimand definition into a concrete study design (data splitting approach).

References:

1. Steyerberg, E. W., & Harrell, F. E. (2016). Prediction models need appropriate internal, internal–external, and external validation. Journal of clinical epidemiology, 69, 245-247.
2. Alpers, R. and Westphal, M. (2023). An estimand framework to guide model and algorithm evaluation in predictive modelling. Manuscript in preparation.
3. Lundberg, I., Johnson, R., & Stewart, B. M. (2021). What is your estimand? Defining the target quantity connects statistical evidence to theory. American Sociological Review, 86(3), 532-565.

*Both authors contributed equally and are listed in alphabetical order.

Tuesday, 05/Sept/2023 2:40pm - 3:00pm
ID: 246 / S30: 3
Presentation Submissions - Regular Session (Default)
Topic of Submission: Time-to-Event Analysis
Keywords: time-to-event analysis, risk, number needed to harm, treatment effect, hazard ratio

A non-parametric proportional risk model to assess a treatment effect in an application to long-term carcinogenicity assays

Lucia Ameis[1], Oliver Kuß[2], Annika Hoyer[3], Kathrin Möllenhoff[1]
[1]Heinrich Heine University Düsseldorf, Germany; [2]German Diabetes Center, Leibniz Institute for Diabetes Research at Heinrich Heine University Düsseldorf, Institute for Biometrics and Epidemiology, Düsseldorf, Germany; [3]Biostatistics and Medical Biometry, Medical School OWL, Bielefeld University, Bielefeld, Germany; lif34pot@uni-duesseldorf.de

Time-to-event analysis often relies on prior parametric assumptions or, if a non-parametric approach was chosen, Cox's proportional hazards model that is inherently tied to an assumption of proportional hazards. This limits the quality of the results in case of any violation of these assumptions. Especially the assumption of proportional hazards was recently criticized for being rarely verified. In addition, most interpretations focus on the hazard ratio, that is often misinterpreted as the relative risk and comes with the restriction of being a conditional measure. Our approach introduces an alternative to the proportional hazard assumption and allows for a direct estimation of the relative risk as well as the absolute measure of the number needed to harm, therefore provides the possibility of an easy and holistic interpretation.

In this talk, we propose a new non-parametric estimator to assess the relative risk of two groups to experience an event under the assumption that the risk is constant over time, namely the proportional risk assumption. Precisely, we first estimate the respective cumulative distribution functions of both groups by means of the Kaplan-Meier estimator and second combine their ratio at different time points to estimate the mean relative risk. We then combine the result with one of the estimated cumulative distribution functions to assess the number needed to harm. This offers the possibility to interpret the treatment effect solely based on a Kaplan-Meier estimator and offers a flexible alternative to Cox's model if the proportional hazard assumption is violated.

We demonstrate the validity of the approach by means of a simulation study and present an application to mortality data of mice from a study investigating the long-term carcinogenicity of piperonyl butoxide.

Monday, 04/Sept/2023 2:20pm - 2:40pm
ID: 115 / **S12: 2**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Critical cross-disciplinary collaboration in dose optimization in oncology

A Roadmap for Novel Oncology Dose Finding Designs

Revathi Ananthakrishnan, frank shen, rong liu, chunsheng he, zhihong yang
BMS, United States of America; revathi.ananthakrishnan@bms.com

Oncology drugs are generally quite toxic and their toxicity is assumed to increase with dose. However, in many new immuno-oncology drugs, the efficacy cannot be assumed to increase with dose but may plateau with dose. Hence finding an optimal dose for safety and efficacy i.e. a dose of the drug that is not too toxic and is not sub-therapeutic is critical. This is why dose finding studies constitute a very important part of the drug development cycle in oncology. Many early phase oncology dose finding designs have been proposed to determine the dose of the study drug that can be used in later phase trials, resulting in a bewildering array of options. The traditional 3+3 design in oncology is a simple algorithm/rule-based design but several other algorithm-based, model-assisted and model-based frameworks for dose finding have been proposed more recently. We conduct a review of the current popular oncology dose finding frameworks, discuss their benefits and downsides, and compare them. We provide an overview of the options and a clear, concise set of guidelines to consider the dose finding frameworks and choose the design that best fits the needs of the current trial.

Thursday, 07/Sept/2023 9:30am - 9:50am
ID: 292 / S59: 4
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Volume-outcome relationships in health care
Keywords: Hospital surgical volume, volume-outcome association, breast cancer

Relationship between hospital volume and medium-term survival in breast cancer surgical and oncologic treatment in Lombardy -Italy

Anita Andreano[1,2], **Maria Grazia Valsecchi**[2], **Antonio Giampiero Russo**[1]
[1]Epidemiology Unit, Agency for Health Protection of Milan, Milan, Italy; [2]School of Medicine and Surgery, University of Milan-Bicocca, Milan, Italy; aandreano@ats-milano.it

The association between surgical, chemotherapy and radiotherapy hospital volume and intermediate-term survival was evaluated in a registry cohort of 18,938 patients (age 19-85) with epithelial breast cancer, diagnosed between Jan 2014 and Dec 2016, resident and treated in five out of eight Lombardy Agencies for Health Protection (AHP) including 78% of Lombardy residents. Cancer characteristics at diagnosis were retrieved from the AHP registries and information on treatment from hospitalization, outpatient and drug databases. Follow-up was investigated through census at Dec 2019. N=789 patients were excluded because of no registered treatment , and 651 because treatments had been performed outside the study area. Of the 17,498 included patients 95% were surgically treated, 38% received CT and 56% RT. We tried to determine which volume was the best proxy of hospital quality. For this purpose, both previous-year and cumulated 3-year volumes were used, considering both all breast surgeries and specific breast cancer surgeries while, as regards to chemotherapy and radiotherapy, both the total and the specific volume for breast cancer delivered from the hospital. Each patient was assigned, for each treatment, the volume of the facility where it was delivered. Only hospitals with a previous-year specific volume of at least five were analysed: 75 for surgery, 24 for radiotherapy, and 61 for chemotherapy. Previous year specific volumes I and III quartiles were (patient/year): 111-596 for surgery, 99-424 for chemotherapy and 213-641 for radiotherapy. The association between volume and death was then estimated using a Cox model with hospital as a random effect. The Hazard Ratio (HR) of death and its relative p values, the AICs of the linear models, and the AICs and graphical trends of the models with the volume as a spline (natural cubic with 3 to 5 knots) were then compared for the different treatment volumes. The AICs of the models with splines were all smaller than those with the linear variable. Both at unadjusted and adjusted analysis, the association between chemotherapy volume and outcome was not significant. For surgery and radiotherapy, the specific average volume of the previous year was chosen for the adjusted analyses, including the following potential confounders: age, stage, morphology, comorbidity index, educational qualification, grading, emergency diagnosis. Predictive mean matching imputation was performed for stage, grading and educational qualification because of missing data (<25% for all). The models with and without random intercept for the hospital were compared, using the likelihood ratio test, which was highly significant both for the linear volume model and for those with splines ($p < 0.00001$). Therefore, the random effects model with the linear volume was compared with those with the volume inserted as a spline with 3 and 4 nodes using the AIC, calculated on the likelihood after integration of the random effect. The HR of death for every 100 unit of volume increase was 0.977 ($p=0.0004$) for surgery and 0.960 ($p=0.04$) for radiotherapy, the latter excluding stage IV patients. However, the association was not linear and, based on the AIC, the model with 3 knots was chosen for both.

Wednesday, 06/Sept/2023 9:30am - 9:50am
ID: 314 / S48: 4
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical hypothesis testing (covariate adjustment, nonparametric methods, multiple comparisons, …), Preclinical drug development, safety and toxicology
Keywords: preclinical research, replication success, sample size calculation, confirmatory research

May the power be with you? Influence of sample size calculation on replication success.

Collazo Anja, Danziger Meggie
Berlin Institute of Health, Germany; anja.collazo@bih-charite.de

In preclinical replication projects such as the Reproducibility Project: Cancer Biology, the central goal is to solidify evidence with respect to a knowledge claim. To enable reliable conclusions from a replication project, sample size choices are an important yet underrated aspect of study design. First, the choice reflects how well researchers balance resources and potential information gain given ethical constraints inherent to animal experimentation. Second, sample size calculation implicitly presents a decision-criterion on whether a replication study might seem unnecessary or unfeasible. Third, many definitions of replication success are dependent on the precision of effect size estimates which in turn depends on sample size. As there is little established guidance, researchers often resort to the “standard” approach using the original effect size estimate to base their sample size estimation on. It has been shown that this likely results in replication failure due to underpowered studies.

Here, we explore how different conceptual starting points for sample size estimation influence the probability to declare replication success. Building on empirical data of 86 original studies from three preclinical replication projects, we conducted a simulation study contrasting the standard approach to calculate sample size for a replication to three other approaches. One approach employs a smallest effect size of interest (SESOI), the safeguard method uses the lower bound of the confidence interval, and the skeptical p-value is a reverse Bayesian method in which a prior centered around the null is applied to the original effect estimate. The approaches differ in how they incorporate uncertainty of the original study estimate into sample size estimation to increase reliability.

Based on the estimated sample sizes, studies were categorized with regard to whether a replication is feasible, unfeasible (ntotal > 280), or not necessary (ntotal < 4). In the SESOI approach, all 86 experiments were carried forward to replication, whereas with standard approach 78/86, the safeguard 68/86 and skeptical p-value methods 49/86 were selected, respectively.

The standard approach on average advised reducing sample sizes in the replication compared to the original study. In contrast, all other approaches suggested an increase in sample sizes for the replication, in accordance with the goal to increase reliability. In addition, we assessed replication success for each of the approaches. The standard approach fares worst, achieving less than 50% replication success given that a true moderate to large effect is present. While the SESOI and safeguard approaches achieve the highest success rates with over 90%, substantially more animals are needed for the replication effort. The skeptical p-value approach best balances success rates and number of animals invested across the 86 experiments.

Our results reveal that the standard approach to sample size estimation for replication fails to increase reliability and decreases chances to declare replication success compared to all other approaches. We reason that preclinical replication studies are worthwhile only if conducted ethically. This might mean that more animals are needed, however, they are used in studies bound to strengthen evidence robustness rather than wasted in studies bound to be inconclusive.

Monday, 04/Sept/2023 5:10pm - 5:30pm
ID: 331 / **S18: 4**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Real world data and evidence
Keywords: RWE, Biosimilar

Real World Evidence Application in Biosimilar Development

Ramin Arani, Jessie Wang, Sreekanth Gattu, Martina Uttenreuther-Fischer, Arne Ring, Samriddhi Buxy Sinha
Sandoz, United States of America; samriddhi_buxy.sinha@sandoz.com

Real-world evidence (RWE), a key part of integrated evidence framework, plays an increasingly important role in optimizing trial design and supporting the approval of new drug development, especially for orphan and oncology drugs. However, the use of RWE in biosimilar development is still exploratory and lacks clear guidance from health authorities. In biosimilar study design, RWE has potential implications for filling historical knowledge gaps to improve study design efficiency, i.e. improving/validating design parameters such as margins or variability; contextualizing study results in the landscape of marketed biosimilar drugs; determining design parameters when historical data for specific endpoints or populations, etc. are not available. A study was initiated to verify the reliability of RWE use as a source of historical knowledge.

This study used AACR GENIE (American Association for Cancer Research, Tumor Information Exchange for Genomics Evidence) real-world data (RWD) to evaluate the treatment effect of immunotherapy compared with chemotherapy in non-small cell lung cancer (NSCLC) patients. Patients were screened based on several criteria, including age, diagnosis, line of treatment (i.e. first line), setting (i.e. monotherapy), etc., to represent a similar population as used in a pivotal Phase III clinical trial.

Patients in the immunotherapy group were matched with those in the chemotherapy group by age, sex, smoking history, and histology based on propensity scores calculated by logistic regression models. Best responses during treatment based on imaging records and medical records were derived, respectively. Results for differences in treatment rwRR (real world response rates), rwPFS (real world progression-free survival), and rwOS (real world overall survival) were compared with those reported in clinical trial publications.

The GENIE database included 1849 patients as of September 2020. Out of the patients who met inclusion and exclusion criteria, 34 and 189 patients received fist line immunotherapy and chemotherapy, respectively. The treatment duration (months) was comparable between immunotherapy versus chemotherapy (Median [Q1, Q3]: 4.17 [2.79, 8.71] vs. 3.06 [1.91, 4.04]). The rwRR based on overall imaging documentation was 35.5% (11/31) for immunotherapy and 19.4% (6/31) for chemotherapy. The median rwPFS was 5.6 month (95% CI: 4.0 – 27.9) versus 8.9 month (95% CI: 5.6 – NA) for immunotherapy versus chemotherapy. The median rwOS was 26.3 months (95% CI: 19.7 – NA) with immunotherapy versus 19.6 months (95% CI: 11.5 – 33.9) with chemotherapy.

This study explored a method to assess the feasibility of applying RWE in biosimilar development. Similar search by involving additional real-world database like Flatiron and/or ConcertAI has also been initiated to provide more insight into the applicability of RWE for designing biosimilar trials. This part of results will be available during presentation.

Tuesday, 05/Sept/2023 5:10pm - 5:30pm
ID: 206 / **S36: 4**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Clinical trial designs (adaptive and platform designs, external controls, …), Statistical hypothesis testing (covariate adjustment, nonparametric methods, multiple comparisons, …)
Keywords: Biomarker, Subgroup, Type I error, Interim analysis

Adaptive enrichment designs for clinical trials with multiple endpoints

Koko Asakura[1], Toshimitsu Hamasaki[2], Frank Bretz[3]
[1]National Cerebral and Cardiovascular Center, Japan; [2]The George Washington University Biostatistics Center, USA; [3]Novartis Pharma AG, Switzerland; koko.a@ncvc.go.jp

We discuss clinical trials that allow adaptive enrichment with prespecified subgroups at an interim analysis and assessment of treatment effect in the enriched subgroups with hypotheses related to two primary endpoints. This setting leads to various methodological, such as: (1) different subgroups may be selected at an interim analysis for each endpoint; (2) depending on the flexibility of the designs, more than one source of multiplicity need be considered due to multiple endpoints, subgroups, and analyses.

In this presentation, we consider a two-stage design where a test intervention is compared with a control intervention with possible adaptations based on conditional power after Stage 1, to enrich to common subgroups for both endpoints. Multiple hypotheses are assessed by following the closure principle and combination tests are used for combining the stagewise p-values. The implications of these designs on power and sample size under the Type I error control are discussed. We illustrate the approaches with an example.

Tuesday, 05/Sept/2023 3:00pm - 3:20pm
ID: 315 / S30: 4
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), Real world data and evidence, Time-to-Event Analysis
Keywords: component-wise gradient boosting, forest health, integrated backward selection, smooth additive Cox model, spatial frailty mode

Modelling tree survival for investigating climate change effects

Nicole H Augustin
University of Edinburgh, United Kingdom; nicole.augustin@ed.ac.uk

Using German forest health monitoring data we investigate the main drivers leading to tree mortality and the association between defoliation and mortality; in particular (a) whether defoliation is a proxy for other covariates (climate, soil, water budget); (b) whether defoliation is a tree response that mitigates the effects of climate change and (c) whether there is a threshold of defoliation which could be used as an early warning sign for irreversible damage.

Results show that environmental drivers leading to tree mortality differ by species, but some are always required in the model. The defoliation effect on mortality differs by species but it is always strong and monotonic. There is some evidence that a defoliation threshold exists for spruce, fir and beech.

We model tree survival with a smooth additive Cox model allowing for random effects taking care of dependence between neighbouring trees and non-linear functions of spatial time varying and functional predictors on defoliation, climate, soil and hydrology characteristics.

Due to the large sample size and large number of parameters, we use parallel computing combined with marginal discretization of covariates. We propose a 'boost forward penalise backward' model selection scheme based on combining component-wise gradient boosting with integrated backward selection.

This is joint work with Axel Albrecht, Heike Puhlmann, Stefan Meining (Forstwissenschaftliche Versuchsanstalt, Freiburg, Germany), Karim Anaya-Izquierdo, Alice Davis (University of Bath), Simon Wood (University of Edinburgh).

Thursday, 07/Sept/2023 11:00am - 11:20am
ID: 119 / **S66: 2**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Advanced survival analysis
Keywords: Causal inference, Frailty models, Survival analysis

A sensitivity analysis approach for the causal hazard ratio in randomized and observational studies

Rachel Axelrod, **Daniel Nevo**
Tel Aviv University, Israel; axelrod1@mail.tau.ac.il

The Hazard Ratio (HR) is often reported as the main causal effect when studying survival data. Despite its popularity, the HR suffers from an unclear causal interpretation due to a built-in selection bias. While alternative approaches exist, the HR remains the most popular measure used by practitioners, and therefore, analysis approaches directly targeting a causally interpretable HR are of interest. A recently proposed alternative is the causal HR, defined as the ratio between hazards across treatment groups among the study participants that would have survived regardless of the assigned study group. We discuss the challenge in identifying the causal HR from the observed data and present a sensitivity analysis approach for identification in randomized controlled trials utilizing a working frailty model. We further extend our framework to adjust for potential confounders using inverse probability of treatment weighting. We present a Cox-based and non-parametric kernel-based estimation under right censoring. We study the finite-sample properties of the proposed estimation methods through simulations and illustrate the utility of our framework using two real-data examples.

Monday, 04/Sept/2023 11:40am - 12:00pm
ID: 159 / **S5: 3**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Finding the right dose – Project Optimus and beyond
Keywords: Dose-Finding, Late-Onset Toxicities, Late-Onset Activity, Phase I Trials, Model-Based

Dose Finding Studies for Therapies with Late-Onset Safety and Activity Outcomes

Helen Barnett[1], Dimitris Kontos[2], Oliver Boix[2], Thomas Jaki[3]
[1]Lancaster University; [2]Bayer; [3]University of Regensburg and University of Cambridge, Germany; thomas.jaki@protonmail.com

In Phase I/II dose-finding trials, the objective is to find the Optimal Biological Dose (OBD), a dose that is both safe and shows sufficient activity that maximises some optimality criterion based on safety and activity. In cancer treatment is typically given over several cycles complicating the identification of the OBD as both toxicity and activity outcomes may occur at any point throughout the follow up of multiple cycles. In this work we present and assess the Joint TITE-CRM, a model-based design for late onset toxicities and activity based on the well-known TITE-CRM. It is found to be superior to both currently available alternative designs that account for late onset bivariate outcomes; a model-assisted method and a bivariate survival design, as well as being both intuitive and computationally feasible.

Monday, 04/Sept/2023 4:50pm - 5:10pm
ID: 367 / S20: 3
Presentation Submissions - Regular Session (Default)
Topic of Submission: Clinical trial designs (adaptive and platform designs, external controls, …)
Keywords: stepped wedge, cohort, time effects, endogenous time-varying covariates, bias

Bias through endogenous time-varying covariates in the analysis of cohort stepped-wedge trials: a simulation study

Jale Basten[1], Daniel Claus[1], Katja Ickstadt[2], Nina Timmesfeld[1]
[1]Department of Medical Informatics, Biometry and Epidemiology, Ruhr-University Bochum, Germany; [2]Faculty of Statistics, TU Dortmund University, Germany; basten@amib.ruhr-uni-bochum.de

One of the major advantages of stepped-wedge cluster-randomised trials (SW-CRTs) over cluster-randomised trials in parallel design is that all participating clusters (e.g. practices) receive the intervention, because they all unidirectional crossover from the control to intervention conditions, which is conducive to recruitment rates [1].

Depending on the intervention, two different approaches can be chosen for SW-CRTs: On the one hand, cross-sectional data can be collected, so that different patients are observed in each step; on the other hand, a cohort of patients can be observed across several steps (cohort data) [2].

Due to the staggered design of SW-CRTs, observations collected under the control condition are, on average, from an earlier calendar time than observations collected under the intervention condition. Thus, the stepped wedge design is susceptible to time effects, such as secular trends, e.g. public policies or seasonal fluctuations. In a cohort design, correlation between measurements within a participant are dependent of the timing in which the observations are made. This raises the possibility that responses may vary over time due to secular trends (external time effects), changes in cohort characteristics (internal time effects), as well as because of changes in treatment. Therefore, a model that allows for time effects is essential [3].

In a longitudinal study, fixed effects can be exogenous or endogenous. Examples of exogenous covariates include baseline variables (age, gender, etc.), function of time, and time-varying variables that are not impacted by prior treatment or prior outcome. In contrast, endogenous covariates are impacted by prior treatment or prior outcome, e.g. the frailty of a participant impacts mental health, but prior intervention (e.g. care program) and prior mental health condition may also impact the frailty of a participant [4].

To analyse the intervention effect in SW-CRTs, we use linear mixed-effects (LME) models with two random effects used to account for clustering (within-cluster correlation) and multiple measurements on participants (within-individual correlation). We will compare model specifications with different fixed effects to investigate which model specification yields unbiased intervention effect estimates in spite of external and internal time effects.

If time-varying confounders are exogenous, we already demonstrated with an extent Monte-Carlo simulation that LME models with fixed categorical time effects additional to the fixed effect of intervention and two random effects used to account for the within-cluster and within-individual correlation seem to produce unbiased estimates of the intervention effect in SW-CRTs with closed and open cohort data even if time-varying exogenous confounders or their functional influence on outcome were unmeasured or unknown and if secular trends occurred [5].

In this talk we will present our results of a simulation study extended to endogenous time-varying covariates that influence participants' responses in cohort SW-CRTs. We seek to find a model approach with the best performance in terms of bias for different realistic data scenarios. Both closed and open cohort data will be considered and the results will be compared.

References

[1] Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. Contemp Clin Trials. 2007 Feb;28(2):182-91. doi: 10.1016/j.cct.2006.05.007. Epub 2006 Jul 7. PMID: 16829207.

[2] Feldman HA, McKinlay SM. Cohort versus cross-sectional design in large field trials: precision, sample size, and a unifying model. Stat Med. 1994 Jan 15;13(1):61-78. doi: 0.1002/sim.4780130108. PMID: 9061841.

[3] Hemming K, Taljaard M, McKenzie JE, Hooper R, Copas A, Thompson JA, Dixon-Woods M, Aldcroft A, Doussau A, Grayling M, Kristunas C, Goldstein CE, Campbell MK, Girling A, Eldridge S, Campbell MJ, Lilford RJ, Weijer C, Forbes AB, Grimshaw JM. Reporting of stepped wedge cluster randomised trials: extension of the CONSORT 2010 statement with explanation and elaboration. BMJ. 2018 Nov 9;363:k1614. doi: 10.1136/bmj.k1614. PMID: 30413417; PMCID: PMC6225589.

[4] Qian T, Klasnja P, Murphy SA. Linear mixed models with endogenous covariates: modeling sequential treatment effects with application to a mobile health study. Stat Sci. 2020;35(3):375-390. doi: 10.1214/19-sts720. Epub 2020 Sep 11. PMID: 33132496; PMCID: PMC7596885.

[5] Basten J, Ickstadt K, Timmesfeld N. Bias through time-varying covariates in the analysis of cohort stepped wedge trials: a simulation study. arXiv e-prints 2023: arXiv:2302.11258.

Tuesday, 05/Sept/2023 2:00pm - 2:20pm
ID: 225 / **S29: 1**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Generalized pairwise comparisons
Keywords: multivariate data, nonparametric inference

Multivariate Outcomes and the Need for Generalized Pairwise Comparisons

Arne Bathke
Universität Salzburg, Austria; abathke@gmail.com

We consider different models for multivariate data, trying to impose as few assumptions as possible, in particular also allowing for endpoints that are not all measured on the same scale. Based on these models, different analysis methods are justifiable for inference. Some of them can be considered part of the class of generalized pairwise comparison (GPC) procedures. We discuss advantages and disadvantages of different approaches in terms of statistical performance, robustness, flexibility, and interpretability. Application of the methodology is illustrated with real data examples.

Monday, 04/Sept/2023 4:30pm - 4:50pm
ID: 162 / S18: 2
Presentation Submissions - Regular Session (Default)
Topic of Submission: Clinical trial designs (adaptive and platform designs, external controls, …), Statistical modelling (regression modelling, prediction models, …), Statistical hypothesis testing (covariate adjustment, nonparametric methods, multiple comparisons, …), Preclinical drug development, safety and toxicology
Keywords: drug accumulation, PK linearity, SAD/MAD studies, superposition principle, time dependent PK

Assessment of pharmacokinetic linearity after repeated drug administration

Alexander Bauer[1], Matts Kagedal[2], Martin J Wolfsegger[1]
[1]Baxalta Innovations GmbH, A Takeda Company, Vienna, Austria; [2]Genentech Inc., South San Francisco, California, USA;
alexander.bauer@takeda.com

The prediction of drug concentration time courses after different dosing scenarios is greatly facilitated if the pharmacokinetics (PK) can be assumed linear. The assumption of linear PK thus needs careful evaluation for any new drug in development. Under linear PK, exposure is proportional to dose (linear PK across doses) and exposure at steady state can be predicted from a single dose based on the superposition principle (linear PK over time). While investigation of dose-proportionality is common practice, evaluation of time dependent PK has received less attention in the literature. In particular, the superposition principle can be used to assess whether the observed extent of accumulation after repeated administration is expected under the premise of linear PK. This work emphasizes the importance of the time related aspect of linear PK by introducing the predictability ratio (PR). Linear PK over time can be concluded if PR = 1. Accumulation is higher than expected if PR >1, and lower if PR <1. If PK data from multiple dose cohorts are available, the PR is assessed for each dose cohort and a supportive hypothesis test can be applied to test for potential differences between doses in PR.

Thursday, 07/Sept/2023 8:50am - 9:10am
ID: 196 / **S58: 2**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Clinical trial designs (adaptive and platform designs, external controls, …)
Keywords: basket trials, power priors, information sharing

Computationally Efficient Basket Trial Designs Based on Empirical Bayes Power Prior Methods

Lukas Baumann, **Meinhard Kieser**
Institute of Medical Biometry, University of Heidelberg, Germany; baumann@imbi.uni-heidelberg.de

Basket trials are used when a new treatment is tested simultaneously in several disjoint subgroups. They are mostly applied in oncology trials, where the subgroups comprise patients with different primary tumor sites but a common biomarker. Usually, basket trials are uncontrolled phase II studies that investigate a binary endpoint such as tumor response. Most of the recently proposed designs for such trials utilize Bayesian tools to partly share the information between baskets depending on the similarity in order to increase the power compared to a separate analysis of each subgroup.

A promising and computationally cheap design was proposed by Fujikawa et al. (2020), where the subgroups are at first analyzed individually using a beta-binomial model. Information is then shared by calculating a weighted sum of the beta posterior parameters of the subgroups, where the weights are based on a similarity measure that is computed for the pairwise comparison of all individual posterior distributions.

Fujikawa's design is closely related to the approach of power priors, specifically to methods using empirical Bayes techniques. Power priors were originally proposed to incorporate historical data. For this application Gravestock & Held (2019) showed that when information is borrowed from several historical studies, calculating weights based only on pairwise similarity with the current study is not optimal. Instead, they proposed an approach that incorporates all of the historical studies at once. We adapt the approach by Gravestock & Held for basket trials and consider other ways to extend Fujikawa's design by incorporating the overall heterogeneity between the baskets.

The designs based on power priors are computationally cheap compared to other Bayesian basket trial designs. Exact calculation of the posterior distribution is fast even for a large number of baskets. Furthermore, for some of the designs even an analytic computation of operating characteristics such as type 1 error rate and power is feasible for a moderate number of subgroups. We compare the performance of these designs to other competing basket trial designs.

The performance of the extended designs is comparable to other basket trial designs in terms of the expected number of correct decisions. In scenarios where only some of the baskets are active, the adaptations improve Fujikawa's design in terms of a smaller type 1 error inflation.

Tuesday, 05/Sept/2023 5:10pm - 5:30pm
ID: 130 / **S38: 3**
Presentation Submissions - Invited Session
Invited Sessions: From multivariate to high-dimensional and functional data
Keywords: Efron's Bootstrap, Factorial Designs, Heteroscedasticity, Multivariate Analysis of Variance, Nonparametric Inference, Quantile-Based Analysis

Quantile-based MANOVA: A new tool for inferring multivariate data in factorial designs

Marléne Baumeister[1], Marc Ditzhaus[2], Markus Pauly[1]
[1]TU Dortmund University, Germany; [2]Otto von Guericke University Magdeburg, Germany; marlene.baumeister@tu-dortmund.de

In various fields, e.g., biology, ecology, medicine, or psychology, several outcome variables are of simultaneous interest leading to multivariate data. For example, an ecologist may study the aggression against predators and the relative reproductive success (fitness) of birds grouped by sex and colour morph. Other examples are psychological tests or different medical quantities, e.g., heart rate, blood pressure, weight, or height of a patient. As pointed out by Warne (2014) , multivariate analysis-of-variance (MANOVA) is "one of the most common multivariate statistical procedures in the social science literature". However, classical MANOVA relies on restrictive assumptions as normality and homogeneity of covariances. But the "normality assumption becomes quasi impossible to justify when moving from univariate to multivariate observations" (Konietschke et al., 2015) and, similarly, homogeneity is often implausible. To overcome these difficulties there are less restrictive mean-based MANOVA concepts proposed for testing global hypothesis about multivariate expectations, e.g Friedrich and Pauly (2018). In case of outliers or distributions with larger tails, however, non-robust estimators like the mean can have some drawbacks. Despite the usage of quantiles is intuitive in that case and often applied in descriptive statistics, e.g. boxplots, quantiles "[appear] to be quite underused in medical research" (Beyerlein, 2014).

Therefore we developed a flexible quantile-based MANOVA method. The approach is adaptable to general factorial designs and has the advantage that it fits to median and other quantile-based statistical methods. To achieve this, we considered two quadratic-form type test statistics and three different strategies for estimating the covariance. The test statistics' distribution is approximated via resampling. We prove that our method is valid in theory and even works in case of general heterogeneous or heteroscedastic data beyond normality. In a simulation study, we compare the novel procedures with state-of-the-art mean-based approaches and observe that the quantile-based approach produces more powerful tests in the case of heavy-tailed data. As an illustrative example we consider heavy-tailed data about the colour morphs of common buzzard chicks.

References

Beyerlein, A. (2014). Quantile Regression–Opportunities and Challenges From a User's Perspective. American Journal of Epidemiology, 180(3):330–331.

Friedrich, S. and Pauly, M. (2018). MATS: Inference for potentially singular and heteroscedastic MANOVA. Journal of Multivariate Analysis, 165:166–179.

Konietschke, F., Bathke, A. C., Harrar, S. W., and Pauly, M. (2015). Parametric and nonparametric bootstrap methods for general MANOVA. Journal of Multivariate Analysis, 140:291–301.

Warne, R. (2014). A Primer on Multivariate Analysis of Variance (MANOVA) for Behavioral Scientists. Practical Assessment, Research & Evaluation, 19:17.

Wednesday, 06/Sept/2023 8:50am - 9:10am
ID: 234 / **S47: 2**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Statistical software engineering in the pharmaceutical industry: Increasing productivity, transparency, and reproducibility via open source collaboration

Refactoring and extending an existing R package across companies - learnings from the crmPack team

Clara Beck[1], Daniel Sabanes Bove[2], Robert Adams[1], Burak Kürsad Günhan[3], Oliver Boix[1], John Kirkpatrick[2], Wojciech Wójciak[2], Marlene Schulte-Goebel[3], Dimitris Kontos[1]
[1]Bayer AG, Germany; [2]Roche; [3]Merck Healthcare KGaA, Germany; clara.beck@bayer.com

Working collaboratively across companies on an open-source project does not come naturally to many statisticians, statistical analysts and developers in the pharmaceutical industry. Although there has been an increasingly strong trend towards more knowledge sharing in recent years, the day-to-day work of many colleagues still consists of finding home-grown, customized solutions for problems and new statistical methods with proprietary software.

Making crmPack an open-source collaboration project was an expedient next step when the need for a flexible approach for the simulation of Phase I dose escalation models came up in multiple companies in parallel. The compelling advantage of this package is its flexible framework that allows easy extension and enhancement of existing methods.

A cross-company team, consisting of pharmaceutical companies, academic institutes, and clinical research organizations, learned to collaborate utilizing agile principles quickly. However, such a collaboration can be very diverse regarding educational background, expertise, and expectations. Initially, it might be challenging to establish a common understanding of how to contribute and a way forward, still guaranteeing a certain level of code quality and set-up reliable communication channels. We would like to share some learnings and best-practices we developed while we have been working together on the extension of crmPack.

Initiatives like crmPack offer the possibility to reach industry-wide standards in the future which will enhance our work, join forces, and use the crowd intelligence to improve quality of code and saving efforts simultaneously while following pharmaceutical industry´s GxP.

Monday, 04/Sept/2023 11:20am - 11:40am
ID: 371 / **S7: 2**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical hypothesis testing (covariate adjustment, nonparametric methods, multiple comparisons, …), Free Contributions
Keywords: nonparametric statistics, nonparametric hypothesis testing, resampling, simultaneous confidence intervals

Almost Omnibus Nonparametric Inference for Two Independent Samples

Jonas Beck[1], **Patrick B. Langthaler**[2], **Arne C. Bathke**[1]
[1]University of Salzburg, Austria; [2]Paracelsus Medical University; jonas.beck@plus.ac.at

Different statistical functionals are omnipresent in nonparametric statistics. Functionals like the Mann-Whitney functional (relative effect) P(X<Y) measure a location or stochastic tendency effect for two independent samples. We extend this by additionally incorporating an overlap index (nonparametric dispersion measure). We develop the joint asymptotic distribution of their rank-based estimators and construct confidence regions based on a resampling approach.

Extending the two-sample rank sum test (Wilcoxon, Mann and Whitney), we propose a new test based on these functionals for the hypothesis of distribution equality. As we simultaneously test different functionals we get a much larger consistency region as in the classical test and in most cases a substantial improvement in the power to the Kolmogorow Smirnov Test, as our simulation shows. We additionally show the usability of our approach by applying the test to different real data sets.

Thursday, 07/Sept/2023 10:40am - 11:20am
ID: 111 / **S65: 1**
Presentation Submissions - Invited Session
Invited Sessions: Advancing clinical trial design in rare diseases
Keywords: rare diseases, development efficiency, external controls, informational designs, basket trials

Innovations in Clinical Development in Rare Diseases in Children and Adolescents

Robert Beckman[1,2], Zoran Antonijevic[2,3], Mercedeh Ghadessi[2,4], Heng Xu[2,5], Cong Chen[2,6], Yi Liu[2,5], Rui Tang[2,7]
[1]Georgetown University Medical Center, United States of America; [2]Drug Information Association Innovative Design Scientific Working Group, United States of America; [3]Abond CRO, United States of America; [4]Bayer Pharmaceuticals, United States of America; [5]Nektar Therapeutics, United States of America; [6]Merck and Co.,Inc., United States of America; [7]Servier Pharmaceuticals, United States of America; rab302@georgetown.edu

Many of the afflictions of children are rare diseases. This creates numerous drug development challenges related to small populations, including limited information about the disease state, enrollment challenges, and diminished incentives for pediatric development of novel therapies by pharmaceutical and biotechnology sponsors. This presentation reviews selected innovations in clinical development that may partially mitigate some of these difficulties, starting with the concept of development efficiency for individual clinical trials, clinical programs (involving multiple trials for a single drug), and clinical portfolios of multiple drugs, and decision analysis as a tool to optimize efficiency. Development efficiency is defined as the ability to reach equally rigorous or more rigorous conclusions in less time, with fewer trial participants, or with fewer resources. We go on to discuss efficient methods for matching targeted therapies to biomarker defined subgroups, methods for eliminating or reducing the need for natural history data to guide rare disease development, the use of basket trials to enhance efficiency by grouping multiple similar disease applications in a single clinical trial, and the use of alternative data sources including historical controls to augment or replace concurrent controls in clinical studies. Greater understanding and broader application of these methods could lead to improved therapies and/or more widespread and rapid access to novel therapies for rare diseases in both children and adults.

Wednesday, 06/Sept/2023 9:50am - 10:10am
ID: 433 / **S49: 5**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Time-to-Event Analysis
Keywords: Cox proportional hazards model, Firth correction, adverse events, early benefit assessment

Evaluation of adverse events in early benefit assessment (Part I): Firth correction for Cox models in the case of zero events

Lars Beckmann, Guido Skipka, Anke Schulz
IQWiG, Cologne, Germany; lars.beckmann@iqwig.de
Note: A companion presentation to this contribution will be given at the "68. GMDS-JAHRESTAGUNG 2023".

This is the first part of a tandem presentation on the topic of the evaluation of adverse events by the Cox proportional hazards regression in early benefit assessment. The second part will be presented at the GMDS Annual Conference in Heilbronn, Germany, 2023.

For the early benefit assessment of drugs in Germany, the pharmaceutical company must describe the extent of an added benefit of the drug to be assessed compared with an appropriate comparator therapy [1]. The confidence interval of a significant effect must lie completely outside a certain corridor around the null effect for the extent of the effect to be regarded as minor, considerable or major. The corridors are defined by different thresholds depending on outcome category (e.g. all-cause mortality, quality of life or adverse events). For endpoints in the category of adverse events, frequently no events in one of the arms are observed, and while the log rank test provides appropriate p-values, the standard Cox proportional hazard regression does not provide valid effect estimates with corresponding confidence intervals. Thus, in the case of a statistically significant effect according to the p-value, the extent of the effect cannot be determined. Consequently, the overall assessment of the early benefit might be hampered.

Heinze and Schemper proposed an adaption of the Firth correction to reduce bias from maximum likelihood estimation for the Cox proportional hazard [2].

To assess the applicability of this approach, we performed a simulation study of time to event analyses with zero events. We will present results from this study and discuss the situations, in which the application of the Firth correction provides reliable estimates that can be used for the assessment of the extent of an added benefit.

In the second part of the tandem presentation, we will discuss the current procedure for time-to-event analysis with zero events and the possible contribution of the Firth correction in the early benefit assessment of adverse events by the IQWiG.

References:

1. IQWiG. General Methods 6.1 [online]. 2022. URL: https://www.iqwig.de/en/about-us/methods/methods-paper/
2. Heinze and Schemper (2001). Biometrics 57(1):114–119.

Wednesday, 06/Sept/2023 11:40am - 12:00pm
ID: 238 / S52: 4
Presentation Submissions - Regular Session (Default)
Topic of Submission: Machine Learning and Data Science, Time-to-Event Analysis
Keywords: Random Survival Forest, Competing Events

Random Survival Forests for Competing Events: A Subdistribution-Based Approach

Charlotte Behning[1], Alexander Bigerl[2], Marvin Wright[3], Moritz Berger[1], Matthias C. Schmid[1]
[1]Institute of Medical Biometry, Informatics and Epidemiology, University Hospital Bonn; [2]DICE Group, Department of Computer Science, Paderborn University, Paderborn, Germany; [3]Leibniz Institute for Prevention Research and Epidemiology - BIPS, Bremen, Germany; behning@imbie.uni-bonn.de

Random Survival Forests (RSF) can be applied to many time-to-event research questions, and are particularly useful in situations where the relationship between the independent variables and the event of interest is rather complex. However, in many clinical settings, the occurrence of the event of interest is affected by competing events, which means that a patient can experience an outcome other than the event of interest. Neglecting the competing event (i.e. regarding competing events as censoring) will typically result in biased estimates of the cumulative incidence function (CIF). A popular approach for dealing with competing events is Fine & Gray's subdistribution hazard model, which performs estimation of the CIF by fitting a single-event model defined on a subdistribution time scale. Here, we integrate concepts from the subdistribution hazard modeling approach into the RSF: We utilize the central feature of RSF - the creation of multiple decision trees, each of which is trained on a random subset of the data. In each tree, the competing event time is replaced by an imputed, possibly right-censored subdistribution time and split rules for single-event RSF are applied. The predictions from the individual trees are then combined to obtain a final prediction. The performance of our proposed method is illustrated by a simulation study.

Tuesday, 05/Sept/2023 5:30pm - 5:50pm
ID: 185 / **S42: 5**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Time-to-Event Analysis
Keywords: non-proportional hazards, delayed treatment effect

Simulation study to compare methods to analyze time-to-event endpoints in trials with delayed treatment effects

Rouven Behnisch, Marietta Kirchner, Meinhard Kieser
Institute of Medical Biometry, University of Heidelberg, Germany; behnisch@imbi.uni-heidelberg.de

The advance of immuno-oncology therapies comes with the challenge to deal with the unique mechanisms of action of these drugs especially when the primary efficacy endpoint is a time-to-event endpoint. It has been shown that the most powerful methods to compare time-to-event endpoints are weighted log-rank tests with weights proportional to the hazard ratio. This implies that the standard log-rank test, often required by regulatory authorities, is most powerful under proportional-hazards alternatives. This assumption is often violated by the mechanism of action leading to delayed treatment effects or crossing of survival curves resulting in a substantial loss in power. Hence, a rather long follow-up period is required to detect a significant effect in immune-oncology trials when the log-rank test is used. Another way to compensate this loss in power would be to prespecify weights proportional to the hazard ratio but is often not feasible since the exact mechanism is usually not known in advance. Recently, different alternatives have been advocated, including well-known methods such as the family of Fleming-Harrington weighted log-rank statistics, accelerated failure time models or additive hazard models, but also newly developed methods such as the modestly weighted log-rank test, the MaxCombo test and tests based on the restricted mean survival time or combinations of different test statistics.

For a better overview over the multitude of methods that have been proposed so far, we have conducted a systematic literature search. The resulting set of methods was then compared systematically with regard to type I error and power in an extensive simulation study. To incorporate different mechanisms of action, we simulate data based on a generalized linear lag model for varying times of study duration, accrual and delay as well as different treatment effects. For methods where parameters need to be prespecified, the influence of misspecification of these parameters on the power was also assessed.

Most of the methods control type I error and achieve reasonable power in case of proportional hazards. A delayed treatment effect results in a power reduction for all methods, but the extent of this reduction varies between methods.

As expected the performance of the log-rank test decreases with increasing treatment delay and there is no single method that performs best in all scenarios so the choice of the optimal analysis strategy depends on the assumed delay pattern.

Monday, 04/Sept/2023 12:20pm - 12:40pm
ID: 406 / **S3: 5**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), Machine Learning and Data Science, Real world data and evidence, Time-to-Event Analysis
Keywords: small data, data pooling, similarity, propensity scores, prediction

Similarity as a basis for data pooling - Improving Local Prediction Models Using External Data

Max Behrens[1], Maryam Farhadizadeh[1], Astrid Pechmann[3], Janbernd Kirschner[3], Angelika Rohde[2], Daniela Zöller[1]
[1]Institute of Medical Biometry and Statistics, University of Freiburg, Germany; [2]University of Freiburg, Department of Mathematical Stochastics, Freiburg im Breisgau, Germany; [3]Department of Neuropediatrics and Muscle Disorders, Faculty of Medicine, Medical Center – University of Freiburg; max.behrens@uniklinik-freiburg.de

Combining data from different sites can provide a larger and more diverse dataset for analysis, which can lead to improved prediction models. For example, when studying a rare disease, a single site may not have enough patients to build a reliable prediction model. However, differences in patient care, case mix, and other factors can pose significant challenges for including data from other data sites. Specifically, this heterogeneity may introduce bias and result in a decreased prediction performance for the target site population when not addressed properly. Further challenges arise when the sample sizes are small, making approaches with a high number of parameters unsuitable. For example, we consider data from the SMArtCARE registry on patients diagnosed with the rare genetic disease spinal muscular atrophy (SMA). Treatment and disease progress evaluation include physiotherapy, which highly depends on the data site, requiring site-specific prediction models for the time to reach a mobility milestone in SMA patients or the mobility score at a specific time point. To address this problem, we propose to quantify the similarity between the target and the external site and to employ this information for including external sites in a weighted manner.

Specifically, we propose to estimate the probability of an individual belonging to the target site using pairwise logistic regression models and to use this probability to assign a higher weight to external individuals similar to the target site individuals than less similar external individuals when building the prediction model. This process is repeated for all pairwise comparisons between the target site and each of the external sites. To incorporate multiple external sites, we standardize the weights across all of them. Since we use weights, this approach can be easily applied to different types of outcomes and prediction models.

In addition to demonstrating the approach using the SMArtCARE registry data, we will evaluate our proposed method using an extensive simulation study, also comparing it to classical approaches like mixed models and regression models with interactions. We demonstrate that the proposed method can overcome the challenges posed by heterogeneity between sites in multi-site data settings and improve the prediction performance of models for a target site. Our approach quantifies similarity between sites using logistic regression and incorporates this information to include external data when building prediction models.

Monday, 04/Sept/2023 11:00am - 11:20am
ID: 263 / **S6: 1**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Biomarkers and diagnostics, Statistical modelling (regression modelling, prediction models, …)
Keywords: Ratio outcomes, Gamma distribution, Frank's copula, Correlated random variables, Dementia research

Modeling the Ratio of Gamma Distributed Random Variables using Frank's Copula

Moritz Berger[1], **Nadja Klein**[2], **Matthias Schmid**[1]
[1]Institute of Medical Biometry, Informatics and Epidemiology, Medical Faculty, University of Bonn; [2]Chair of Uncertainty Quantification and Statistical Learning, Research Center for Trustworthy Data Science and Security (UA Ruhr) and Department of Statistics (Technical University Dortmund); moritz.berger@imbie.uni-bonn.de

In clinical and epidemiological studies one frequently encounters the ratio of two possibly correlated components. Typical examples are, among others, the LDL/HDL cholesterol ratio in cardiovascular research, the CD4/CD8 ratio in HIV research and the GEFC/REFC ratio in fundus autofluorescence imaging. In regression analysis with a ratio outcome, a reasonable assumption is that the two components follow a gamma distribution each, thereby accounting for the positivity of the component values and the skewness of their distributions. If independence between the two components can be assumed, the ratio of two gamma distributed variables follows a generalized beta distribution of the second kind (GB2; Kleiber and Kotz, 2003). Several regression approaches for the GB2 distribution have been proposed recently. For positively correlated components, Berger et al. (2019) developed a regression model based on Kibble's bivariate gamma distribution, where one of the parameters is directly interpretable in terms of the Pearson correlation coefficient between the two components. Regarding the ratio of two negatively correlated components no regression modeling strategy exists so far.

To address this issue, we propose a regression model where the joint bivariate distribution of the two gamma distributed random variables is given by Frank's copula (Genest, 1987). The model explicitly accounts for a negative (or positive) correlation between the two components. It also allows for different forms of the two marginal distributions with possibly unequal rate and shape parameters. The probability density function of the ratio conditional on covariate values and distributional parameters of interest can be derived in a very flexible way. We illustrate the approach analyzing data from dementia research, where cerebrospinal fluid biomarkers are used for early diagnoses of Alzheimer's disease. In this application, measurements of the amyloid-beta 42 protein and total tau protein exhibit a clearly negative correlation.

References:

1. M. Berger, M. Wagner, and M. Schmid. Modeling biomarker ratios with gamma distributed components. *The Annals of Applied Statistics*, 13:548–572, 2019.
2. C. Genest. Frank's family of bivariate distributions. *Biometrika*, 74:549–555, 1987.
3. C. Kleiber and S. Kotz. *Statistical Size Distributions in Economics and Actuarial Sciences*. Wiley, Hoboken, 2003.

Thursday, 07/Sept/2023 11:40am - 12:00pm
ID: 370 / S69: 4
Presentation Submissions - Regular Session (Default)
Topic of Submission: Machine Learning and Data Science, High dimensional data, genetic and x-omics data
Keywords: DNA sequencing, DRAGEN, High-Throughput Sequencing, Illumina NovaSeq 6000, Next Generation Sequencing

Pre-processing and quality control of whole genome sequencing data: a case study using 9000 samples from the GENESIS-HD study

Raphael O. Betschart[1], Domingo Aguilera-Garcia[2], Hugo Babel[1], Stefan Blankenberg[1,3,4], Linlin Guo[3], Holger Moch[2], Dagmar Seidl[2], Felix Thalén[1], Alexandre Thiéry[1], Raphael Twerenbold[3,4], Tanja Zeller[3,4], Martin Zoche[2], Andreas Ziegler[1,3,5,6]

[1]Cardio-CARE, Medizincampus Davos, Switzerland; [2]Institute of Pathology and Molecular Pathology, University Hospital Zurich, Zurich, Switzerland; [3]University Center of Cardiovascular Science and Department of Cardiology, University Heart and Vascular Center, University Medical Center Eppendorf, Hamburg, Germany; [4]German Center for Cardiovascular Science (DZHK); partner site Hamburg/Kiel/Lübeck, Hamburg, Germany; [5]Swiss Institute of Bioinformatics, Lausanne, Switzerland; [6]School Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Pietermaritzburg, South Africa;
andreas.ziegler@cardio-care.ch

Rapid advances in high-throughput DNA sequencing technologies have enabled the conduct of large-scale whole genome sequencing (WGS) studies. Before association analysis between phenotypes and genotyped can be conducted, extensive pre-processing and quality control (QC) of the raw sequence data need to be performed. This case study describes the pre-processing pipeline and QC framework we have selected for the GENEtic SequencIng Study Hamburg-Davos (GENESIS-HD), a study involving more than 9000 human whole genomes. All samples were sequenced on a single Illumina NovaSeq 6000 with an average coverage of 35x, using a PCR-free protocol and unique dual indices (UDI). For quality control (QC), one genome in a bottle (GIAB) trio was sequenced in tetraplicate, and one GIAB sample was successfully sequenced 70 times in different runs. In this presentation, we illustrate the application of important QC metrics to the data at the different pre-processing stages. We provide empirical data for the compression of raw data using the novel original read archive (ORA). Our results show that the most important quality metrics for sample filtering were ancestry, sample cross-contamination, deviations from the expected Het/Hom ratio, relatedness, and too low coverage. The compression ratio of the raw files using ORA was 5:1, and the compression time was linear with respect to genome coverage. In summary, the pre-processing, joint calling, and QC of large WGS studies is feasible in reasonable time, and efficient QC procedures are readily available.

Tuesday, 05/Sept/2023 4:50pm - 5:10pm
ID: 209 / **S36: 3**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Clinical trial designs (adaptive and platform designs, external controls, …), Statistical hypothesis testing (covariate adjustment, nonparametric methods, multiple comparisons, …)
Keywords: Adaptive design, endpoint selection, sample size reassessment, composite endpoint

Adaptive selection of binary composite endpoints and sample size reassessment based on blinded data

Marta Bofill Roig[1], Guadalupe Gómez Melis[2], Martin Posch[1], Franz Koenig[1]
[1]Medical University of Vienna, Austria; [2]Universitat Politècnica de Catalunya; marta.bofillroig@meduniwien.ac.at

For randomized clinical trials where a single, primary, binary endpoint would require unfeasibly large sample sizes, composite endpoints (CEs) are widely chosen as the primary endpoint. Despite being commonly used, CEs entail challenges in designing and interpreting results. Given that the components may be of different relevance and have different effect sizes, the choice of components must be made carefully. Especially, sample size calculations for composite binary endpoints depend not only on the anticipated effect sizes and event probabilities of the composite components but also on the correlation between them. However, information on the correlation between endpoints is usually not reported in the literature, which can be an obstacle to designing future sound trials.

We consider two-arm randomized controlled trials with a primary composite binary endpoint and an endpoint that consists only of the clinically more important component of the CE. We propose a trial design that allows an adaptive modification of the primary endpoint based on blinded information obtained at an interim analysis. Especially, we consider a decision rule to select between a CE and its most relevant component as the primary endpoint. The decision rule chooses the endpoint with the lower estimated required sample size. Additionally, the sample size is reassessed using the estimated event probabilities and correlation, and the expected effect sizes of the composite components. We investigate the statistical power and significance level under the proposed design through simulations. We show that the adaptive design is equally or more powerful than designs without adaptive modification on the primary endpoint. Besides, the targeted power is achieved even if the correlation is misspecified at the planning stage while maintaining the type 1 error. Extensions to trials with multiple composite components of different relevance and with more than two arms are also presented. Finally, we illustrate the proposal by means of a cardiology trial using the *eselect* R package, which provides the implementation of the proposed design.

Wednesday, 06/Sept/2023 8:50am - 9:10am
ID: 194 / S48: 2
Presentation Submissions - Regular Session (Default)
Topic of Submission: Clinical trial designs (adaptive and platform designs, external controls, …)
Keywords: sample size recalculation, adaptive study design, three-stage clinical trials

Sample size recalculation in three-stage clinical trials

Björn Bokelmann[1], Geraldine Rauch[1], Jan Meis[2], Meinhard Kieser[2], Carolin Herrmann[1]
[1]Institute of Biometry and Clinical Epidemiology, Charité - Universitätsmedizin Berlin; [2]Institute of Medical Biometry, University Medical Center Ruprechts-Karls University Heidelberg; bjoern.bokelmann@charite.de

Choosing an adequate sample size for a clinical trial is an important task and can be challenging. To find a sample size, which puts as few as possible patients at risk while maintaining a high power, among others, information about the effect size of the medical treatment is required. Wrong assumptions about the effect size could lead to underpowering or oversizing of a trial. One potential remedy to this problem are multi-stage trials combined with sample size recalculation. In a first stage, a number of patients is recruited and the outcomes of the primary endpoint are examined in an interim analysis. The obtained information from the interim analysis is then used to (re-)calculate the sample size of the following stage. There exists research about sample size recalculation for two-stage trials and the according approaches are already applied in practice (Friede & Kieser, 2006; Bauer et al. 2016). However, for designs with more than two stages, previous literature only examines the potential of sample size reduction due to efficacy and futility stopping, without considering the possibility of sample size recalculation (Chen, 1997; Chen, 2008). In our research, we consider three stage trials, with the option for futility and efficacy stopping after the first two stages. We examine under which conditions and to what extent sample size recalculation at the final stage could prevent underpowering or oversizing, if the assumed effect size deviates from the true effect size. While sample size recalculation could yield these potential benefits, it also poses an application challenge due to the uncertainty about the total number of patients to recruit when starting the trial. To measure this disadvantage, we also examine the variance of the sample size.

References

1. Friede, T., & Kieser, M. (2006). Sample size recalculation in internal pilot study designs: a review. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, *48*(4), 537-555.
2. Bauer, P., Bretz, F., Dragalin, V., König, F., & Wassmer, G. (2016). Twenty-five years of confirmatory adaptive designs: opportunities and pitfalls. *Statistics in Medicine*, *35*(3), 325-347.
3. Chen, T. T. (1997). Optimal three-stage designs for phase II cancer clinical trials. *Statistics in medicine*, *16*(23), 2701-2711.
4. Chen, K., & Shan, M. (2008). Optimal and minimax three-stage designs for phase II oncology clinical trials. *Contemporary Clinical Trials*, *29*(1), 32-41.

Tuesday, 05/Sept/2023 2:40pm - 3:00pm
ID: 261 / S35: 3
Presentation Submissions - Regular Session (Default)
Topic of Submission: Epidemiology
Keywords: parametric g-formula, IDEFICS/I.Family cohort, childhood obesity, causal inference, observational data

Estimating the effects of hypothetical behavioral interventions on overweight/obesity incidence using observational data: Methodological challenges and practical considerations

Claudia Börnhorst[1], Iris Pigeot[1,2], Stefaan De Henauw[3], Annarita Formisano[4], Lauren Lissner[5], Denéz Molnár[6], Luis A Morena[7,8], Michael Tornaritis[9], Toomas Veidebaum[10], Tanja Vrijkotte[11], Maike Wolters[1], Vanessa Didelez[1,2], on behalf of the GrowH! consortium[1]

[1]Leibniz Institute for Prevention Research and Epidemiology - BIPS, Bremen, Germany; [2]Institute of Statistics, Faculty of Mathematics and Computer Science, University of Bremen, Bremen, Germany; [3]Department of Public Health and Primary Care, Ghent University, Ghent, Belgium; [4]Institute of Food Sciences, National Research Council, Avellino, Italy; [5]School of Public Health and Community Medicine, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden; [6]Department of Pediatrics, Medical School, University of Pécs, Pécs, Hungary; [7]GENUD (Growth, Exercise, Nutrition and Development) Research Group, Faculty of Health Sciences, Universidad de Zaragoza, Instituto Agroalimentario de Aragón (IA2), Instituto de Investigación Sanitaria Aragón (IIS Aragón), Zaragoza, Spain; [8]Centro de Investigación Biomédica en Red de Fisiopatología de la Obesidad y Nutrición (CIBERObn), Madrid, Spain; [9]Research and Education Institute of Child Health, Strovolos, Cyprus; [10]National Institute for Health Development, Estonian Centre of Behavioral and Health Sciences, Tallinn, Estonia; [11]Department of Public and Occupational Health Amsterdam UMC, Amsterdam, The Netherlands; boern@leibniz-bips.de

Description: Although randomized studies previously assessed short-term effects of behavioral factors on overweight/obesity (OW/OB), they could not assess the effects under real life conditions nor over long time spans. Using methods of causal inference, we therefore aimed to study the long-term effects of hypothetical behavioral interventions on OW/OB from childhood to adolescence, i.e. to answer questions such as "What would happen to the incidence of OW/OB if all children continuously adhered to screen time recommendations over a period of 13 years". Our sample comprised 10 877 children aged 2 to <10 years at baseline who participated in the well-phenotyped IDEFICS/I.Family cohort[1]. Children were followed over 13 years from 2007/2008 to 2020/21. The risk of developing OW/OB was estimated under various single and joint hypothetical behavioral interventions using the parametric g-formula[2].

The 13-year risk of developing OW/OB was found to be 30.7 [28.4;32.7] percent under no intervention and 25.4 [22.1;27.7] percent when multiple interventions were imposed jointly, corresponding to a risk reduction of 17%. The most effective interventions were to meet screen time recommendations and to meet moderate-to-vigorous physical activity recommendations which could reduce the incidence of OW/OB by -2.2 [-4.4;-0.7] and -2.1 [-3.7;-0.8] percentage points (risk difference [confidence interval]), respectively. Meeting sleep recommendations (-0.6 [-1.1;-0.3]) showed a similar intervention effect as compared to increasing sleep duration by 30 minutes/day (-0.6 [-0.9;-0.3]). If all children were members in a sports club, the incidence of OW/OB could be reduced by -1.6 [-2.7;-0.4] percentage points over a 13-year period. Unexpectedly, reducing the consumption of sugar-sweetened beverages by 1 drink/day increased the risk by 0.73 [0.16;1.4] percentage points; however, this effect disappeared in a sensitivity analysis where exposures were redefined so as to precede the outcome by several years.

Our analysis is one of only few practical applications of the parametric g-formula to cohort data. We will discuss the plausibility of the underlying assumptions and point to practical challenges in its implementation in the context of our specific research question: These address, for instance, the irregular and fairly long time intervals between waves and the modelling of time-varying covariates.

To conclude, we evaluate the utility of our approach for estimating intervention effects based on observational data and highlight potential sources of bias. We make suggestions for strengthening confidence in any results obtained from observational data by following the principle of target trial emulation.

References

1. Ahrens W, Siani A, Adan R, et al. Cohort Profile: The transition from childhood to adolescence in European children-how I.Family extends the IDEFICS cohort. Int J Epidemiol 2017;46(5):1394-1395j.
2. Naimi AI, Cole SR, Kennedy EH. An introduction to g methods. Int J Epidemiol 2017;46(2):756-762.

Wednesday, 06/Sept/2023 9:10am - 9:30am
ID: 373 / **S43: 3**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Estimands and causal inference, Epidemiology, Real world data and evidence

Target trial emulation avoids bias due to non-alignment at time-zero in studies on site-specific effectiveness of screening colonoscopy

Malte Braitmaier[1], Sarina Schwarz[1], Vanessa Didelez[1,2], Ulrike Haug[1,3]
[1]Leibniz Institute for Prevention Research and Epidemiology - BIPS, Germany; [2]University of Bremen, Faculty of Mathematics and Computer Science, Germany; [3]University of Bremen, Faculty of Human and Health Sciences, Germany;
braitmaier@leibniz-bips.de

Objective: Observational studies often suffer from biases due to flaws in the study design. We illustrate the target trial emulation (TTE) framework as a principle to avoid such self-inflicted biases using the example of investigating differences in site-specific effectiveness of colonoscopy screening. Particularly, previous observational studies reported a higher effectiveness of colonoscopy in preventing distal vs. proximal colorectal cancer. We aim to assess whether this difference arises from design-induced biases. Furthermore, we give a structural explanation of biases resulting from flawed study designs.

Methods: We used the same dataset underlying our recently published analysis (Braitmaier et al. 2022) based on German claims data (20% population coverage). We investigated the site-specific effectiveness of colonoscopy screening over 11 years of follow-up in 55-69-year-old persons using a target trial emulation framework to avoid self-inflicted biases. To explain the discrepancy between our findings and the published literature, we re-analyzed the same data, but using a naïve study design that is commonly found in observational studies, where exposure assignment is based on pre-baseline information thus violating time-zero-alignment. Finally, we analyzed the resulting bias using causal diagrams.

Results: While in the recently published analysis using TTE, the relative risk (RR) of distal and proximal CRC in the screening colonoscopy vs. control group was similar (RR: 0.67 for distal, 0.70 for proximal), the analysis with “time zero violation” indicated a difference in site-specific performance. In this latter analysis, the RR was 0.41 for distal CRC and 0.66 for proximal CRC.

Conclusions: The various potential sources of bias in the analysis of observational (‘real-world’) data, many of which avoidable, are still not widely understood. By contrasting a principled TTE with a standard but naïve study design, we demonstrated the potential for bias. We explain this as a type of collider-stratification bias due to pre-baseline selection. The designs of observational studies should emulate the key design elements from randomized trials, e.g., time-zero-alignment, to avoid such biases.

References:

Braitmaier M, Schwarz S, Kollhorst B, Senore C, Didelez V, Haug U (2022) Screening colonoscopy similarly prevented distal and proximal colorectal cancer: a prospective study among 55-69-year-olds; Journal of Clinical Epidemiology; 149: 118-126

Tuesday, 05/Sept/2023 11:00am - 11:20am
ID: 355 / S27: 1
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Statistical issues in health care provider comparisons
Keywords: registry, arthroplasty, hip, knee, reporting

Comparing implants, hospitals and surgeons: lessons learned from the Swiss National Hip & Knee Joint Registry SIRIS

Christian Brand[1], Martin Beck[2], Lilianna Bolliger[1], Bernhard Christen[3], Vilijam Zdravkovic[4]
[1]Universität Bern, Switzerland; [2]Orthopädische Klinik Luzern; [3]Articon, Spezialpraxis für Gelenkchirurgie; [4]Kantonsspital St. Gallen; christian.brand@ispm.unibe.ch

The Swiss joint registry SIRIS has been registering hip and knee replacements since 2012. It currently achieves 100% participation from eligible hospitals and captures approx. 98% of eligible procedures (subject to patient consent). It primarily evaluates implants (hip stem-cup combinations, hemi-arthroplasty stem-head combinations, total and partial knee systems) against national averages of reference categories. The primary outcome is all cause revision rates at different time points with special prominence given to 2-year 'early' revision rates. It also evaluates hospitals and surgeons against national averages, primarily drawing on Kaplan-Meier graphs and funnel plots for visualisation purposes. In all of these activities statistical outliers are detected, reported and subject to certain protocols in order to facilitate quality improvement.

The principal problems encountered are small volume units, clustering of observations (strong local effects and dependency of results on individual units), camouflaging of effects, moral hazard dilemmas caused by the potential of "gaming" among participants, incomplete data, time dependency of effects due to changing circumstances and evolving registry quality. As is the case with other international arthroplasty registries, the general approach to analysis and reporting could be characterised as big data driven, relying on SRS confidence intervals, stratification by type of prosthesis, limited risk-adjustment to account for patient mix, accounting for censoring events, and an underlying assumption that other more complex issues such as clustering of observations will to some extent blur away as numbers get bigger. Analyses for reporting purposes often have to make pragmatic choices in order to enable timely reporting of results, provide information that recipients can actually understand and even avoid technical restrictions such as excessive computing demands. Technically "more correct" analyses are therefore typically undertaken for special tasks or general verification purposes, but not routinely reported.

This presentation will provide an overview of the registry's activities and highlight the challenges encountered on specific examples. In particular, it is noted that different reporting levels require different solutions and considerations. Hospitals and surgeons are typically evaluated on selected standard procedures and from certain time periods which are then risk-adjusted and presented as funnel plots. Outliers are identified in the conventional way (>99.8% limit). Implants, on the other hand, already form typical design groups which are approved for particular medical conditions and therefore exhibit a fair degree of homogeneity regarding patient mix. SIRIS uses a simpler, more descriptive approach in the evaluation of grouped implants by simply defining fixed acceptability thresholds for revision rates against which all implants with sufficient group sizes are evaluated. Outliers are those that exceed those thresholds with a clear margin (with a further distinction between possible and definitive outlier).

Tuesday, 05/Sept/2023 4:50pm - 5:10pm
ID: 222 / **S37: 3**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Causal estimands for time to event data
Keywords: time to event, estimand, regulatory, causal

Estimands for time to event data: a regulator's view

Andreas Brandt
BfArM (Federal Institute for Drugs and Medical Devices), Germany; andreas.brandt@bfarm.de

The publication of the ICH E9 addendum on estimands and sensitivity analysis moved the importance of the estimand as the exactly defined target of inference into the focus of regulatory scientific advice and assessment of clinical trials. For time-to-event endpoints, awareness for the need to account for post-randomisation events such as discontinuation or change of treatment existed before the ICH E9 addendum and is reflected in regulatory recommendations for study design and censoring rules. However, the estimand framework draws attention to the need for a precise definition of the underlying question prior to aligning design and analysis accordingly, and to the need to differentiate between intercurrent events and missing data. As it appears questionable that all estimands of interest can be appropriately addressed by the still most commonly used approach of adapting censoring rules, alternatives may need to be advocated. However, regulatory experience with alternatives is still limited.

The hazard ratio (HR) estimated based on the Cox model is still the most commonly used summary measure for the treatment effect in studies with a time-to-event endpoint. However, it has been increasingly criticized due to its lack of causal interpretation. Furthermore, deviations from non-proportional hazards in certain therapeutic areas have aroused additional interest in alternative effect measures. From a regulatory point of view, an effect measure should not only inform benefit-risk decisions for approval but provide relevant information for labelling across the patient population that can be easily interpreted by patients and prescribers. The HR has never been the only effect measure to support regulatory decisions and labelling, but several measures are used to provide the overall picture. Still, stronger emphasis on less commonly used alternative summary measures allowing a causal interpretation may be useful.

Tuesday, 05/Sept/2023 11:40am - 12:00pm
ID: 445 / S26: 3
Presentation Submissions - Regular Session (Default)
Topic of Submission: Clinical trial designs (adaptive and platform designs, external controls, …), Statistical hypothesis testing (covariate adjustment, nonparametric methods, multiple comparisons, …)

Control of essential type I error rates in clinical trials with multiple hypotheses

Werner Brannath[1], **Frank Bretz**[2]
[1]University Bremen, Germany; [2]Novartis AG, Switzerland; brannath@uni-bremen.de

The talk is about alternatives to the control of the familywise error rate in clinical trials with multiple hypotheses. The focus will be on concepts that control type I error rates only so far, as they are relevant to patients outside and after the trial. Focusing on studies with multiple populations, the familywise expected loss (FWEL; Maurer et al., 2022) and the population-wise error rate (PWER; Brannath et al., 2022) will be introduced as examples. Furthermore, focusing on multi-arm and platform trials with the possibility of dropping treatments mid-trial, it will be discussed how one could account for a mid-trial reduction of the post-trial risks when dropping a treatment. The solutions will be motivated with independent clinical trials, for which no multiplicity adjustment is required. The talk will end with a discussion and outlook on further questions and future research concerning the control of essential type I error rates.

Literature

1. Brannath, W., Hillner, C., Kornelius, R. (2023). The population-wise error rate for clinical trials with overlapping populations. Statistical Methods in Medical Research, 32(2):334-352
2. Maurer, W., Bretz, F., and Xun, X. (2023). Optimal test procedures for multiple hypotheses controlling the familywise expected loss. Biometrics, to appear.
3. Brannath, W. (2023). Discussion on "Optimal test procedures for multiple hypotheses controlling the familywise expected loss" by Willi Maurer, Frank Bretz, and Xiaolei Xun. Biometrics, to appear.

Thursday, 07/Sept/2023 10:40am - 11:00am
ID: 154 / **S70: 1**
Presentation Submissions - Regular Session (Default)
Topic of Submission: COVID pandemic, Epidemiology
Keywords: reproductive number, neutral methods comparison, COVID-19, real-time estimation

Why are different estimates of the effective reproductive number so different? A case study on COVID-19 in Germany

Elisabeth Brockhaus[1], **Johannes Bracher**[1,2]
[1]Karlsruhe Institute of Technology, Germany; [2]Heidelberg Institute for Theoretical Studies; johannes.bracher@kit.edu

The effective reproductive number has taken a central role in the scientific, political and public discussion during the COVID-19 pandemic, with numerous real-time estimates of this quantity routinely published. Disagreement between these estimates can be substantial, and may lead to confusion among decision makers and the general public. In this work we compare different estimates of the effective reproductive number of COVID-19 in Germany during the time period from October 2020 through September 2021. We consider agreement between estimates from the same method but published at different time points (within-method agreement) as well as retrospective agreement across different approaches (between-method agreement). The former is based on an archive of real-time estimates compiled from public repositories of various academic groups. While for some approaches, estimates are very stable over time and hardly subject to revisions, others display considerable fluctuations. To assess between-method agreement, we reproduced the estimates generated by different groups using a variety of statistical approaches, standardizing analytical choices in order to assess how they contribute to the observed disagreement. These analytical choices include the data source, data pre-processing steps, assumed generation time distribution, statistical tuning parameters and temporal alignment of estimates. We find these user choices to be at least as important as the choice of statistical method among the growing number of available options. They should thus be communicated transparently along with published estimates.

Thursday, 07/Sept/2023 11:20am - 11:40am
ID: 320 / **S69: 3**
Presentation Submissions - Regular Session (Default)
Topic of Submission: High dimensional data, genetic and x-omics data
Keywords: polygenic scores, missing data, imputation, proxy SNPs

The impact of missing SNPs in the calculation of polygenic scores

Hanna C. B. Brudermann, Inke R. König
Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Campus Lübeck, Lübeck, Germany;
ha.brudermann@uni-luebeck.de

Polygenic scores (PGS) aggregate the information of many genome-wide markers – mostly single nucleotide polymorphisms (SNPs) – to estimate the genetic susceptibility of a person to a specific phenotype. Over the few last years, guidelines on how to construct PGS have been published (Choi et al. 2020), and the Polygenic Score Catalog (Lambert et al. 2021) is a free resource to screen and download PGS.

When applying a previously published PGS to new data often not all markers that are part of the PGS are available, and those which differ in their quality. If, for example, quality-controlled genotyped and imputed data are used, SNPs fall into different categories of availability: high or low quality and directly genotyped or imputed, or unavailable.

However, knowledge of the impact of various types and degrees of missing markers on the performance of available PGS is limited.

In 2018, Chagnon et al. investigated the influence of missing markers on the calculation of a PGS. They compared the gold standard score based on all genotypes with scores resulting from two different strategies. In the first, the missing markers are omitted in the calculation of the score; and in the second, the missing genotypes are replaced with the genotypes of a proxy SNP with a predefined linkage disequilibrium. The resulting scores were compared with the gold standard regarding correlation and AUC among other PGS quality measures. The results showed that the use of a proxy SNP is generally better than omitting the marker but that attention has to be paid if the missing marker has a relatively high effect size.

In this work, a comprehensive simulation study is performed to extend the methods of Chagnon et al. (2018) in a practically important way. Given that mostly imputed data is typically used, we now consider not only genotyped but also imputed SNPs of different imputation quality. Therefore, for a specific missing marker, one can theoretically choose between a proxy SNP and an imputed one. Nevertheless, it is common for many markers to be missing again after post-imputation quality control.

The first results show that the results of Chagnon et al. (2018) hold for those SNPs that are still missing after imputation. Also, SNPs with a high info score (< 0.9) after imputation show similar behavior as very good proxy SNPs ($r^2 \geq 0.9$) while imputed SNPs with an info score < 0.1 still behave equally to good proxy SNPs ($0.6 \leq r^2 < 0.8$) for low frequencies of missing genotypes (< 20%); and worse than good proxy SNPs for higher frequencies of missing genotypes.

Using both of Chagnon's et al. (2018) strategies to work around missing markers and combine these with the usage of imputed markers, the impact of different degrees of missing markers is investigated. From this, a guideline can be derived for the practical use of PGS.

Literature:

1. Lambert et al. 2021 Nat Genet 53:420-5; doi: 10.1038/s41588-021-00783-5.
2. Choi et al. 2020 Nat Prot 15:2759-72; doi: 10.1038/s41596-020-0353-1.
3. Chagnon et al. 2018 PLoS One 13(7); doi: 10.1371/journal.pone.0200630.

Tuesday, 05/Sept/2023 11:20am - 11:40am
ID: 459 / **S25: 2**
Presentation Submissions - Featured Session
Featured Sessions: Young Statisticians Sessions and Panel Discussion

A two-step approach for analysing time-to-event data under non-proportional hazards

Jonas Elias Brugger, **Franz König**
Medical University of Vienna, Austria; jonas.brugger@meduniwien.ac.at

Survival analysis is a statistical method for evaluating time-to-event data, such as death or disease progression, in oncology trials. Usually, in such oncology trials, a new treatment regimen is compared to a control group, normally the standard of care. Traditional survival analysis methods like the Cox-proportional hazards model or the log-rank test assume the hazard ratio of two groups to be constant over time. However, this assumption is often violated in real-world applications. An example of that are immuno-oncology drugs, which often exhibit a delayed onset of their effects. To address this, more robust methods for survival analysis under non-proportional hazards have been developed. We propose a two-step procedure for comparing hazard functions of two groups in the presence of non-proportional hazards. The procedure starts with a pre-test to assess the proportional hazards assumption, followed by a method for comparing hazard functions that is conditioned on the pre-test result. In a simple framework, depending on the pre-test results either a standard log-rank test or a weighted log-rank test will be performed. We show for which scenarios such a two-step procedure might yield a type 1 error rate inflation and discuss how strict control can be achieved. The efficacy of the two-step approach will be evaluated through comparison with established methods such as weighted log-rank tests or max-combo test in broad simulation study.

Tuesday, 05/Sept/2023 11:20am - 11:40am
ID: 377 / **S24: 2**
Presentation Submissions - Regular Session (Default)
Topic of Submission: High dimensional data, genetic and x-omics data
Keywords: deep learning, dimension reduction, variable selection, differentiable programming, single-cell RNA-sequencing data

Combining Boosting with Neural Networks for Structuring Latent Representations of Single-Cell RNA-Sequencing Data

Niklas Brunn, Maren Hackenberg, Harald Binder
Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Stefan-Meier-Straße 26, 79106, Freiburg, Germany; niklas.brunn@uniklinik-freiburg.de

Dimension reduction is an important step in the analysis of single-cell RNA-sequencing (scRNA-seq) data for identifying underlying patterns. The corresponding low-dimensional representation should have properties such as disentangled dimensions, i.e., different dimensions correspond to distinct underlying factors of variation, and interpretability, e.g., by identifying a small set of characteristic genes for each dimension. For structuring the representation of scRNA-seq data accordingly, we propose to combine feature selection based on componentwise boosting with neural networks for dimension reduction. More precisely, we use an autoencoder architecture that is implicitly regularized by componentwise boosting when minimizing the reconstruction loss. There, componentwise boosting allows capturing a small number of explanatory features in each dimension, hence arriving at an interpretable representation. To derive pseudo-targets for the boosting approach, we use constrained versions of the negative gradients of the reconstruction loss w.r.t. the different components of the current representation. Specifically, a constraint ensures that, for a given dimension, only features are selected that are complementary to the information already encoded in the other dimensions, thus resulting in disentangled dimensions. We use differentiable programming for differentiating through the boosting step in the joint optimization of the boosting component and the neural networks. For illustration, we apply our approach to scRNA-seq data from cortical neurons of mice. The results show that we can identify a small subset of genes for each dimension that characterizes distinct cell types. We furthermore illustrate how our approach can be extended to incorporate temporal development patterns, such as cellular differentiation programs.

Tuesday, 05/Sept/2023 11:00am - 11:20am
ID: 241 / **S22: 1**
Presentation Submissions - Invited Session
Invited Sessions: Net benefit, win odds, and win ratio: Methods, analysis, and interpretation
Keywords: Rank Procedures, Net Treatment Benefit, Win Ratio, Success Odds, Win Odds

Net Benefit, Success Odds, and Win Ratio for Non-Censored Observations

Edgar Brunner
University Medical Center, Germany; ebrunne1@gwdg.de

Nonparametric endpoints for clinical trials based on the Mann-Whitney effect $\theta^*=P(X<Y)$ are discussed in the recent literature in statistical models involving censored observations as well as not involving censored observations but allowing for ties. The last aspect shall be the focus of this talk. Although stated in several papers in statistical journals that the Mann-Whitney effect θ^*, (or $\theta=\theta^*+½\ P(X=Y)$ if ties are involved) is an easy and intuitive treatment effect, this was not perceived in clinical medicine. The interpretation of the net treatment benefit NTB = P(X<Y)-P(X>Y), (Buyse, 2010), the win ratio WR = P(X<Y) / P(X>Y), (Pocock et al., 2012, Wang et al., 2016), and the success odds SO = $\theta/(1-\theta)$ (Dong et al., 2019; Brunner et al., 2021), however, has been accepted in clinical medicine, in particular when considering prioritized composite endpoints see, e.g., Redfors, 2020). The statistical properties, advantages, and problems of the three quantities NTB, WR, and SO will be discussed here.

In case of no ties, WR=SO and both quantities can be interpreted as the chance of obtaining a better result under treatment 1 than under treatment 2. In case of ties, WR can no longer be interpreted in this way since the numerator of WR in not the complement of the denominator. For prioritized composite endpoints, NTB can be estimated by generalized pairwise comparisons (GPC). A simple relation of the GPC to rankings (and in turn to WR and SO) enables the application of well-known results from rank methods. Since NTB is a linear transformation of θ, NTB=2θ–1, all results from rank methods can immediately be applied. This is not the case for WR and SO since these are non-linear transformations leading to problems at the margins 0 and 1. Even unbiased estimations of WR and SO are issues. Procedures for testing H_0: $\theta = ½$ (and in turn WR=1 or SO=1) are available from the literature while confidence intervals for larger values of NTB, WR, or SO are issues, in particular if the samples sizes are not very large.

References

1. Brunner, E. et al. (2021). Win odds: an adaptation of the win ratio to include ties. Statistics in Medicine 40, 3367--3384.
2. Buyse, M. (2010) Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. Statistics in Medicine 29, 3245--3257.
3. Dong, G. et al. (2020). The Win Ratio: On Interpretation and Handling of Ties. Statistics in Biopharmaceutical Research 12, 99--106.
4. Pocock, S.J. et al. (2012). The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. European heart journal 33, 176--182.
5. Redfors, B. et al. (2020). The win ratio approach for composite endpoints: practical guidance based on previous experience. European Heart Journal 41, 4391--4399.
6. Wang, D. et al. (2016). A win ratio approach to comparing continuous non-normal outcomes in clinical trials. Pharmaceutical Statistics 15, 238--245.

Wednesday, 06/Sept/2023 9:10am - 9:30am
ID: 384 / **S45: 3**
Presentation Submissions - Regular Session (Default)
Topic of Submission: High dimensional data, genetic and x-omics data
Keywords: Variable importance, Random forest, High-dimensional data, Permutation

Deriving interpretable thresholds for Variable Importance in Random Forests by permutation

Hannes Buchner[1], Laura Schlieker[1], Maria Blanco[1], Tim Mueller[1], Armin Ott[2], Roman Hornung[3,4]
[1]Staburo GmbH, Aschauer Str. 26a, 81549 München, Germany; [2]Roche Diagnostics GmbH, MMDHA, Nonnenwald 2, 82377 Penzberg, Germany; [3]Institute for Medical Information Processing, Biometry and Epidemiology, University of Munich; [4]Munich Center for Machine Learning (MCML); buchner@staburo.de

In the context of clinical research and in particular precision medicine the identification of predictive or prognostic biomarkers is of utmost importance. Especially when dealing with high-dimensional data discriminating between informative and uninformative variables plays a crucial role. Machine Learning approaches and especially Random Forests are promising approaches in this situation as the variable importance of a Random Forest can serve as a decision guidance for the identification of potentially relevant variables.

Many different approaches for Random Forest variable importance have been proposed and evaluated (e.g., Degenhardt et al. 2019, Speiser et al. 2019). One of the algorithms is the well-performing Boruta method (Kursa and Rudnicki 2010), which adds permutated - and thus uninformative - versions of each variable (so-called shadow variables) to the set of predictors.

We propose a variation of the Boruta method, which is independent of the simulations runs and which compares the variable importance of each covariate directly with the permuted version of the covariate. In addition, in this method, the uninformative versions are generated by permutating the rows of the dataset, which preserves the relationship between the original variables. We aim to evaluate the relevance of the variables based on different criteria, e.g., proportion of positive difference in paired VIMP, mean of the shadow VIMPs and distance between paired distributions.

We examine our method on real data sets of varying sizes and compare its performance to the Boruta algorithm.

ID: 470 / **Plenary 3: 1**
Presentation Submissions - Featured Session
Featured Sessions: Keynote
Keywords: Causal-inspired machine learning, Computational physiology, Domain Adaptation, Electronic health records, Empirical Bayes

Learning from other Intensive Care Units: can we improve statistical predictions?

Peter Bühlmann
ETH Zurich, Switzerland; buehlmann@stat.math.ethz.ch

We discuss the problem of predicting individual patient status in intensive care. While there is massive amount of data available from many medical centers, their integration for a particular intensive care unit or individual patient is challenging. We describe conceptual modeling paradigms for generalization and domain adaptation to new units, combining Empirical Bayes methods and Causal-Inspired Machine Learning. Empirical validations of such approaches provide some interesting insight.

Thursday, 07/Sept/2023 9:50am - 10:10am
ID: 318 / **S60: 5**
Presentation Submissions - Regular Session (Default)
Topic of Submission: High dimensional data, genetic and x-omics data, Time-to-Event Analysis
Keywords: penalized regression, variable selection, high-dimensional data, survival analysis, competing risks

High-Dimensional Variable Selection for Competing Risks with Cooperative Penalized Regression

Lukas Burk[1,2,3], **Andreas Bender**[2,3], **Marvin N. Wright**[1]
[1]Leibniz Institute for Prevention Research and Epidemiology - BIPS, Bremen; [2]Department of Statistics, LMU Munich; [3]Munich Center for Machine Learning, LMU Munich; burk@leibniz-bips.de

Variable selection is an important step in the analysis of high-dimensional omics data, yet there are limited options for survival outcomes in the presence of competing risks. Commonly employed penalized Cox regression considers each event type separately through cause-specific models, neglecting possibly shared information between them.

We adapt the feature-weighted elastic net (fwelnet), a generalization of the elastic net algorithm, to survival outcomes and competing risks. For two causes, our proposed algorithm fits two alternating cause-specific fwelnet models, where each model receives the coefficient vector of the complementary model as prior information. We dub this “cooperative penalized regression”, as this approach enables the modeling of competing risk data with cause-specific models while accounting for shared information between causes. Predictors that are shrunken towards zero in the model for the first cause will receive larger penalization weights in the model for the second cause and vice-versa. Through multiple iterations, this process ensures a stronger penalization of uninformative predictors in both models.

We demonstrate our method’s variable selection capabilities on simulated genomics data and real-world bladder cancer microarray data. We evaluate selection performance using the positive predictive value (PPV) and false positive rate (FPR) for the correct selection of informative features. The benchmark compares results with cause-specific penalized Cox regression, random survival forests, and likelihood-boosted Cox regression (CoxBoost). Results show cooperative penalized regression to yield higher PPV and lower FPR in settings where mutual information is present, which indicates that our approach is more effective at selecting informative features, while being less likely to select uninformative features. In settings where no mutual information is present, variable selection performance is similar to cause-specific penalized Cox regression.

Tuesday, 05/Sept/2023 11:20am - 12:00pm
ID: 256 / **S22: 2**
Presentation Submissions - Invited Session
Invited Sessions: Net benefit, win odds, and win ratio: Methods, analysis, and interpretation

Use and interpretation of the net treatment benefit, success odds, and win ratio for censored and non-censored data

Marc Buyse
IDDI, Belgium; marc.buyse@iddi.com

Three measures of treatment effect have been proposed for generalized pairwise comparisons (GPC): the net treatment benefit (NTB), the win ratio (WR) and the success odds (SO). WR has gained much popularity in cardiovascular clinical trials, where the assumption of proportional hazards is not overly restrictive, in which case WR can be interpreted as the reciprocal of the hazard ratio. As a relative measure of treatment effect, WR is likely to be similar across subgroups of patients of different prognosis. However WR ignores ties and will therefore overestimate the treatment effect in the presence of ties (equal outcome values for the two patients of a pair, which may arise from a coarse time scale or the use of a threshold of clinical relevance in the pairwise comparisons). In addition, WR does not have a simple interpretation for outcomes other than times to event (continuous or ordered categorical outcomes). For such outcomes as well as for times to event, an alternative measure of treatment effect is NTB, which is the probability for a random patient in the treatment group to do better than a random patient in the control group, minus the probability of the opposite situation. The NTB is an absolute measure of treatment effect, and as such it is likely to vary across subgroups of patients of different prognosis. However an absolute measure of effect is required to combine outcomes that capture treatment benefits and harms, and NTB has the advantage that the contributions of prioritized outcomes to the overall NTB are additive. SO is a simple transformation of NTB. It is equal to WR in the absence of ties, but SO does account for ties otherwise. The pros and cons of all three measures will be illustrated using actual trials in oncology and cardiology.

Thursday, 07/Sept/2023 8:30am - 8:50am
ID: 361 / **S63: 1**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), High dimensional data, genetic and x-omics data, Statistical hypothesis testing (covariate adjustment, nonparametric methods, multiple comparisons, …)

DESpace: a novel analysis framework to discover spatially variable genes

Peiying Cai[1], Mark D Robinson[1], Simone Tiberi[1,2]
[1]University of Zurich, Switzerland; [2]University of Bologna, Italy; simone.tiberi@unibo.it

Background

Spatially resolved transcriptomics (SRT) technologies allow measuring gene expression profiles, while also retaining information of the spatial tissue. SRT technologies have led to the release of novel methods that take advantage of the joint availability of mRNA abundance and spatial information. Notably, several computational tools have been developed to identify spatially variable genes (SVGs), i.e., genes whose expression profiles vary across tissue. Nonetheless, current approaches for SVG detection present some limitations; in particular:

i) most methods are computationally intensive;

ii) biological replicates are not allowed;

iii) information about known spatial structures (usually) cannot be incorporated ;

iv) testing cannot be performed on specific regions of interest (e.g., white matter in brain cortex).

Methodology

We propose *DESpace*, an intuitive framework for identifying SVGs based on differential testing across spatial clusters. These clusters represent spatially neighbouring cells with similar expression profiles, and can be obtained via spatial clustering tools (e.g., BayesSpace, StLearn, Giotto and PRECAST), or via pathologists' annotations. We use these clusters as a proxy for the actual spatial information. We then employ *edgeR*, a popular tool for differential expression analyses, to perform differential testing across spatial clusters. Intuitively, if the mRNA abundance of a gene is significantly associated to the spatial clusters, then it varies across the tissue, which indicates a SVG.

Clearly, our framework relies on spatial clusters being available and summarizing the main spatial features of the data. Nonetheless, even in the absence of pre-computed annotations, spatially resolved clustering tools allow generating clusters that accurately summarize the spatial structure of gene expression.

Additionally, *DESpace* presents some unique features compared to currently available SVG tools; in fact, our framework:

i) can model multiple samples, reducing the uncertainty that characterizes inference performed from individual samples, and identifying genes with coherent spatial patterns across biological replicates;

ii) allows identifying the key areas of the tissue affected by SVG, testing if the average expression in a particular region of interest (e.g., cancer tissue) is significantly higher or lower than the average expression of the remaining tissue (e.g., non-cancer tissue), hence enabling scientists to investigate changes in mRNA abundance in specific areas which may be of particular interest.

Finally, our method is flexible, and can input any type of SRT data.

Benchmarking

We performed extensive benchmarks of our approach and various competitors (MERINGUE, nnSVG, SpaGCN, SPARK, SPARK-X, SpatialDE, SpatialDE2, and trendsceek). In particular, starting from three real spatial omics datasets as anchor data, we generated various semi-simulated datasets, with a wide variety of spatial patterns. Our approach displays well calibrated false discovery rates, and higher true positive rate than all competitors considered. Furthermore, when analyzing real data, the genes identified by *DESpace* are more coherent across replicates, than those detected by other SVG methods.

Availability

DESpace is implemented as an R package, currently available on GitHub, and is accompanied by an example usage vignettes: https://github.com/peicai/DESpace

DESpace was also submitted to Bioconductor, where it should appear in a few weeks.

A pre-print (in preparation) will follow in the coming weeks.

Tuesday, 05/Sept/2023 4:30pm - 4:50pm
ID: 411 / S41: 2
Presentation Submissions - Regular Session (Default)
Topic of Submission: Clinical trial designs (adaptive and platform designs, external controls, …)

Robust incorporation of external information in two-arm trial hypothesis testing

Silvia Calderazzo, Manuel Wiesenfarth, Annette Kopp-Schneider
German Cancer Research Center, Germany; s.calderazzo@dkfz.de

When designing a clinical trial, external information on the trial's parameters of interest is sometimes available. The Bayesian approach allows borrowing external information through the adoption of informative prior distributions. In two-arm trials, external information may be available separately for the treatment arm and/or the control arm mean. A difficulty in this situation is that the type I error rate under the informative prior analysis typically depends on the (unknown) true common mean under the null hypothesis, and its maximum may even reach 1. We propose a compromise approach to testing which aims at achieving type I error rates intermediate between the no borrowing (i.e., frequentist) and full borrowing approaches, according to a pre-specified weight parameter. While dependence on the true unknown control parameter value leads to an only approximate correspondence between the target and the realized type I error rate compromise, an explicit upper bound on type I error rate can still be enforced. Such an upper bound may be of advantage in a regulatory setting and improve transparency in communicating the trial design. A dynamic method tailored to hypothesis testing is also proposed to adaptively estimate the weight parameter from the observed data. Simulations are performed to show the properties of the approach under various prior-data conflict and prior informativeness configurations.

Monday, 04/Sept/2023 12:15pm - 12:40pm
ID: 483 / **S4: 4**
Presentation Submissions - Featured Session
Featured Sessions: Biometrical Journal Showcase - Editor's Selection
Keywords: complete records; missing data; multiple imputation; sensitivity analysis

Missing data: A statistical framework for practice

James Robert Carpenter[1,2]
[1]London School of Hygiene & Tropical Medicine, United Kingdom; [2]MRC Clinical Trials Unit at UCL, UK;
james.carpenter@lshtm.ac.uk

Missing data are ubiquitous in medical research, yet there is still uncertainty among many analysts over when restricting to the complete records is likely to be acceptable, when more complex methods (e.g. maximum likelihood, multiple imputation and Bayesian methods) should be used, how they relate to each other and the role of sensitivity analysis.

Based on Carpenter and Smuk (2021), this talk seeks to equip practitioners to address the issues mentioned above by presenting a framework for analysis of partially observed data and illustrative examples, alongside an overview of how the various missing data methodologies in the literature relate. In particular, we describe how multiple imputation can be readily used for sensitivity analyses, which are still infrequently performed.

The ideas are illustrated with a cohort study, a multi-centre case control study and a randomised clinical trial.

Reference:

Carpenter, JR, Smuk, M. Missing data: A statistical framework for practice. *Biometrical Journal*. 2021; 63: 915– 947. https://doi.org/10.1002/bimj.202000196

Tuesday, 05/Sept/2023 12:20pm - 12:40pm
ID: 291 / S28: 5
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical hypothesis testing (covariate adjustment, nonparametric methods, multiple comparisons, …)
Keywords: N-of-1 trial, single-case study, confidence intervals

Statistical considerations on the coverage probability of a confidence interval when sequentially combining n-of-1 studies in a cumulative meta-analysis

Eleonora Carrozzo[1], Georg Zimmermann[2], Arne C. Bathke[3], Daniel Neunhaeuserer[4], Josef Niebauer[1], Stefan Tino Kulnik[1]

[1]Ludwig Boltzmann Institute for Digital Health and Prevention, Salzburg, Austria; [2]Department of Research and Innovation, Paracelsus Medical University, Salzburg, Austria; [3]Department of Artificial Intelligence and Human Interfaces, University of Salzburg, Austria; [4]Sport and Exercise Medicine Division, Department of Medicine, University of Padova, Italy;
eleonora.carrozzo@dhp.lbg.ac.at

N-of-1 randomised clinical trials are receiving broader attention in healthcare research when assessing the effect of interventions. The conventional method to establish intervention effect is a two-arm randomised controlled trial (RCT), where participants are randomised to receive an experimental intervention or a control condition. In contrast, in an N-of-1 design, the individual acts as their own control condition. N-of-1 trials might lead to higher patient care quality while identifying intervention effectiveness of an intervention at an individual level.

However, N-of-1 implementation still presents methodological issues and barriers, such as a general lack of procedural knowledge.

We previously investigated whether sequentially combining the results of single N-of-1 trials in a random-effects meta-analysis allows us to detect statistically significant intervention effects with fewer participants than in a traditional, prospectively powered two-arm RCT. Using data from a crossover RCT, the results showed that the same statistical inference as in an RCT was reached, but requiring fewer participants.

However, it is well known that performing a meta-analysis under a random-effects model systematically underestimates the nominal confidence level of the confidence interval for the overall effect size. Given the promising previous results, further investigation into the methodological properties of the sequential procedure is needed. In the present study, we performed a simulation study both under the null hypothesis and in power, in order to understand how and when this procedure may best be adopted in practice.

Thursday, 07/Sept/2023 11:40am - 12:00pm
ID: 270 / **S70: 4**
Presentation Submissions - Regular Session (Default)
Topic of Submission: COVID pandemic, Meta-Analysis and Systematic Reviews
Keywords: Covid, Systematic Review, prevention, testing

Systematic review on prevention and testing strategies for COVID-pandemic control in economic comparison

Noah Alessandro Castioni, **Eva Herrmann**
Goethe-Universität Frankfurt am Main, Germany; noah@castioni.eu

Introduction

At the time of writing, the Covid 19 pandemic, prevalent since late 2019, is still having an immense impact on healthcare systems and nations worldwide. In relation to this implementation of a wide variety of preventative measures such as social distancing requirements and mask-wearing guidelines. Due to measures and the medical effects, enormous economic consequences could be observed in many European economies, including the German one (Destatis, 2021). Therefore, health care and policy makers have been consistently confronted with the trade-off between different prevention strategies since the beginning of the pandemic.

The aim of this work was therefore to systematically identify publications that evaluate individual measures or combinations of these from both a medical and economic perspective. In a further step, the findings of the included primary studies were then summarized using defined and widely used synthesis methods to capture and present overarching effects.

Methods

A systematic database search was performed using PubMed and WebofScience. The search terms primarily targeted cost-effectiveness analyses (CEAs), net-benefit analyses and cost-per-death-averted analyses (CDA). The results of the systematic literature search were summarized within a systematic review using the "vote counting" method, using qualitative scoring based on the observed direction of effect. Here, a classification of the prevention measures into the four measure categories masks, testing, hygiene and "public health" took place.

Results

Of the total of over 4,000 screening publications, 21 studies met the established inclusion and exclusion criteria. Because some studies looked at different combinations of measures, a total of 66 scenarios resulted in the synthesis. It should be noted that many of the studies were based on mathematical modelling rather than experimentally collected data. For the net benefit studies, the decision rule was a net benefit greater than 0 as a positive effect direction. For the CEA, an empirically based effectiveness threshold of €74,159 (price level: 2010) per QALY was used. For the remaining CDAs, the assumed QALY-loss of a Covid-19-related death was multiplied by the previous effectiveness threshold.

Of the 66 scenarios, 40 (61%) showed a positive and 26 (39%) a negative direction of effect. When the scenarios were divided into two groups, sorted according to the underlying infection incidence (mostly according to the reproduction factor), it became apparent that the prevention measures were primarily cost-efficient, especially in the case of high infection incidence. Only the measure category "public health" was cost-efficient regardless of the infection incidence.

Discussion

A broad but relatively heterogeneous selection of primary literature was found, regarding the different measures considered as well as their chosen method of analysis. Nevertheless, some suitable studies could be identified, which overall, independent of further factors, primarily stated a cost-effectiveness of the prevention measures against Covid-19. Infection incidence was identified as an important parameter. Specifically, the cost-effectiveness varied primarily based on the assumed reproduction factor. However, a dependence on the economic level was also found. Macroeconomic net benefit studies were majority cost-effective, whereas CEAs presented a split picture. Interestingly, as the pandemic unfolded, assumptions about the underlying infection patterns also changed.

Thursday, 07/Sept/2023 9:50am - 10:10am
ID: 410 / S63: 5
Presentation Submissions - Regular Session (Default)
Topic of Submission: High dimensional data, genetic and x-omics data, Statistical hypothesis testing (covariate adjustment, nonparametric methods, multiple comparisons, …)
Keywords: conditional independence, distance covariance, GWAS, complex disease, high-dimensional data

Testing for associations in genomic data with distances and kernels: From unconditional to conditional settings

Fernando Castro-Prado[1,2], Wenceslao Gonzalez-Manteiga[1], Javier Costas[2], Dominic Edelmann[3]
[1]University of Santiago de Compostela, Spain; [2]Health Research Institute of Santiago de Compostela, Spain; [3]German Cancer Research Centre, Heidelberg, Germany; f.castro.prado@usc.es

Distance covariance is an association measure that characterises general statistical independence (not only the linear one) between random vectors on arbitrary metric spaces (not only Euclidean ones). It is dual to the to the Hilbert–Schmidt independence criterion, popular in the machine learning community. With the toolbox of any of the two schools (i.e., strong negative type distances or characteristic kernels, respectively), it is possible to provide meaningful insight into the analysis of data from genome-wide association studies. We briefly introduce some work of us in which we apply these techniques to the search for genetic variants with significant marginal effects on a phenotypical trait of interest, and to the detection of gene-gene interactions. At this point, we wonder what happens when we try to test for such associations conditioning on an environmental covariate of interest. This yields to the conditional version of distance covariance, adapted to the particular geometry that we define to account for the structure of our genetic data. We show some theoretical properties of the resulting test statistic and we explore the performance of our methodology with simulations and a real data example.

Monday, 04/Sept/2023 5:10pm - 5:30pm
ID: 428 / **S19: 4**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Clinical trial designs (adaptive and platform designs, external controls, …), Estimands and causal inference, Time-to-Event Analysis

Assessment of the treatment effect in dose ranging studies with time to event endpoints accounting for the intercurrent event of dose reductions

Arunava Chakravartty, Zheng Li
Novartis, United States of America; arunava.chakravartty@novartis.com

Dose finding trials in Oncology have traditionally been based on sequential dosing cohorts where escalation or de-escalation is guided by the incidence short term dose limiting toxicities. However in the recent past there has been greater push towards supplementing such sequential dose finding designs by randomized dose cohorts evaluating toxicity, tolerability and key efficacy endpoints in order to better optimize the dose prior to starting the Ph 3 pivotal study

Progression free survival (PFS) is a common clinical endpoint in oncology studies to assess the clinical benefit . When used in such randomized dose optimization studies, PFS data can guide the selection of the dose It is typically analyzed per the intention to treat principle. However, a challenge here would be to compare different doses because of the intercurrent event of dose reduction. As the patients in the higher dose groups reduce their doses during the course of the study it makes the cohorts less distinct. Any comparison between the doses, the treatment effect would be underestimated per ITT principle.

In this talk we present two case studies one in pre-market and other in post market dose finding to explore the implications of such dose reductions. We propose to use the estimand framework to assess the role of such dose reductions as an intercurrent event when assessing the PFS difference and present different strategies to handle it. We have considered three estimators multistate survival model, time dependent survival, and G-estimation to estimate the treatment benefit between different dose groups. The performance of these methods will be evaluated by different simulation scenarios and a real clinical study.

Wednesday, 06/Sept/2023 11:20am - 11:40am
ID: 333 / **S53: 3**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Real world data and evidence, Time-to-Event Analysis, Statistical hypothesis testing (covariate adjustment, nonparametric methods, multiple comparisons, …)
Keywords: survival extrapolation, model selection, health technology assessment

On selecting a parametric model to predict long-term survival to support health technology assessment

X. Gregory Chen, Sajjad Rafiq
Biostatistics and Research Decision Sciences, MSD, Switzerland; stat1013@gmail.com

Long-term extrapolations/predictions from parametric survival models for time-to-event outcomes fitted based on clinical data are routinely used to provide inputs for cost-effectiveness analysis in support of health technology assessment.

For any given dataset, typically, several parametric models are fitted. Some make a general distributional assumption for survival time (typically incl. exponential, Weibull, Gompertz, log-logistic, log-normal, Generalized gamma, Generalized F), while others allow more flexible specification (e.g. natural cubic spline model by Royston-Parmar, two-piecewise parametric model).

There is hence still a crucial task for statisticians after model fitting to select the best predictive model or at least provide adequate evaluation of the prediction performance over all fitted models, given available data and information, some of which may not be used in the fitting. The standard approach in practice for model selection relies on the ranking of some goodness-of-fit criteria (e.g. AIC, BIC) and/or visual assessment of the Kaplan-Meier curve versus the fitted curve.

In this talk/poster,

- Firstly we discuss an alternative procedure to use pair-wise likelihood ratio test to make forward model selection (from the simplest model to more complex ones) that takes into account the interdependency of the 7 common distributional assumptions. The interdependency guides the pairwise comparison (not every pair out of the 7 distributions need to be tested). We will evaluate the performance of the proposed procedure via simulation.
- Secondly, we consider more prediction-focus evaluation criteria for any parametric model, and propose a graphical approach to simultaneously screen the internal and external validity of fitted models, better informing the model selection. The external validity could be evaluated based on RWD from relevant general population, or approximated via cross-validation procedure.

Tuesday, 05/Sept/2023 11:40am - 12:00pm
ID: 467 / **S25: 3**
Presentation Submissions - Featured Session
Featured Sessions: Young Statisticians Sessions and Panel Discussion
Keywords: Sample size recalculation, three-arm, noninferiority trial, skewed outcomes

Sample size recalculation for a skewed outcome in two-stage three-arm sequential noninferiority clinical trials: a simulation study

Maria Vittoria Chiaruttini[1], Danila Azzolina[2], Alessandro Desideri[3], Dario Gregori[1]
[1]Unit of Biostatistics, Epidemiology and Public Health, Department of Cardiac, Thoracic and Vascular Sciences, University of Padova, Padova, Italy; [2]Department of Environmental and Preventive Science, University of Ferrara, Ferrara, Italy;
[3]Cardiovascular Research Foundation, S. Giacomo Hospital, Castelfranco Veneto, Italy; mariavittoria.chiaruttini@ubep.unipd.it

The gold-standard design for non-inferiority studies is recommended by regulatory authorities in the United States as well as in Europe: it consists in a three-arm design including placebo in addition to the experimental and the active comparator groups (Koch A. & Röhmel J.,2013). Three-arm non-inferiority trials are challenging for the hypothesis formulation, and their design is often characterized by uncertainty in estimating the experimental treatment effect. Some methods have been proposed to optimize the recalculation of the sample size at interim analysis for Gaussian, Bernoulli, and Poisson outcomes but not for continuous skewed outcome.

Firstly, our research aims to evaluate, through simulations, how the algorithm for the recalculation of the sample size at interim in three-arm non-inferiority studies, with particular reference to Lei's design (Lei, 2020), is able to provide the acceptable level of power even in the presence of deviation from normality; secondly, to improve the applicability of the resampling algorithm in the study design with a gamma distributed outcome of interest.

After demonstrating that the power is still maintained at the predetermined level even in the presence of a deviation from normality, we questioned the hypothesis of homoskedasticity in the case of asymmetrical outcomes simulated as gamma variables. Thus, we provided a resampling algorithm that keeps the normal assumption but accounts for different standard deviations to be set across the three arms.

We found that if the discrepancy between the group variances is considered, we can estimate an adequate initial sample size to achieve the desired power, avoiding the risk of overestimation (saving patients) or underestimation (saving power), in case of gamma distributed outcomes. We provided a motivating example from COSTAMI trial (Desideri A. et al., 2003).

Lastly, we developed an intuitive Web application for the sample size/coverage probability estimation. The tool helps to keep track of the properties of the design, as it provides an estimate of the probability of success/failure of the study, giving us the possibility to choose the best reliable set of parameters to optimize the resources available for the trial.

Thursday, 07/Sept/2023 10:40am - 11:00am
ID: 279 / S64: 1
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), Estimands and causal inference, Time-to-Event Analysis

Asking the right questions when assessing overall survival in a randomized clinical trial that allows for cross over: practical considerations and a case study in cell therapy

Silvia Colicino, Alessandro Previtali
BMS, Switzerland; Silvia.Colicino@bms.com

Background Treatment switching (TS) occurs in randomized clinical trials when patients discontinue their randomly assigned treatment and start a new therapy. Although ethically and clinically justified, TS presents a difficult problem for statisticians trying to ascertain the causal effects of interventions, particularly when assessing long-term time-to-event endpoints such as overall survival (OS). The difficulty arises as TS may occur after randomization but before observing the variable of interest, e.g., a death occurring after TS when assessing OS. Following the spirit of ICH E9(R1) on estimands, the clinical question of interest should drive how TS is handled in the analysis. In practice, especially when seeking regulatory approval, long-term time-to-event endpoints are assessed using an intention-to-treat (ITT) approach comparing treatments as they were initially randomized, i.e., regardless occurrence of TS. Additional methodologies were developed to assess treatment effect (TE) in the hypothetical scenario in which TS would not have occurred. These methods allow the relative effect of the experimental treatment over the control to be isolated by removing (or reducing) the potential benefit of switching. These methods include the rank preserving structural failure time (RPSFT), the two-stage accelerated failure time (2-AFT) and the inverse probability of censoring weighting (IPCW) models. Focusing on a special case of TS hereby referred to as cross over (CO), in which patients were only allowed to switch from control to experimental treatment, this work aims to i) contextualize these analyses in terms of the clinical question of interest, to ii) clarify their interpretation and role in regulatory submissions and to iii) provide an example of their application using a case-study in cell therapy.

Case-study TRANSFORM (NCT03575351) is a global, phase 3 study comparing lisocabtagene maraleucel (liso-cel) versus standard of care (SOC) as second-line therapy for primary refractory or early relapsed large B-cell lymphoma patients considered eligible for autologous stem cell transplantation. A total of 184 patients were randomized 1:1 to either the liso-cel or the SOC arms. CO was allowed upon confirmation by investigators of pre-defined clinical criteria. OS was one of the key secondary endpoints and was analyzed using the ITT approach while the RPSFT, 2-AFT and IPCW models were presented as pre-specified supportive analyses.

Results When the TE was estimated ignoring TS (i.e., following the ITT approach), the study did not demonstrate an improvement in OS (HR = 0.724; 95% CI: 0.443, 1.183). However, when the TE was estimated assuming CO did not occur, results from the 2-AFT and RPSFT models showed a favourable TE for liso-cel (HR = 0.415; 95% CI: 0.251, 0.686 and HR = 0.279; 95% CI: 0.145, 0.537, respectively). IPCW could not be implemented due to data limitations.

Discussion The definition of clear clinical objectives when evaluating long-term time-to-event endpoints is paramount to decide how TS should be analytically addressed. While regulators may be more inclined to focus on the ITT principle, other agencies such as payers may be also interested in evaluating the relative effect assuming TS did not occur. Together, these approaches contribute to a comprehensive evaluation of efficacy.

Tuesday, 05/Sept/2023 5:30pm - 5:50pm
ID: 300 / S36: 5
Presentation Submissions - Regular Session (Default)
Topic of Submission: Clinical trial designs (adaptive and platform designs, external controls, …)
Keywords: R package, Bayesian adaptive design, CARA, RAR

A flexible simulation framework of Bayesian adaptive designs

Dominique-Laurent Couturier[1], Elizabeth G. Ryan[2], Thomas Jaki[1,3], Stephane Heritier[2]
[1]MRC Biostatistics Unit, University of Cambridge, United Kingdom; [2]School of Public Health and Preventive Medicine, Monash University, Australia; [3]Chair for Computational Statistics, Faculty of Informatics and Data Science, University of Regensburg, Germany; dlc48@cam.ac.uk

The growth of Bayesian adaptive designs has been hampered by the lack of software readily available to statisticians. Part of the problem is due to the burden generated by Monte Carlo Markov Chains (MCMC) typically used to compute posterior distributions.

In this work, we follow a different approach based on the Laplace approximation to circumvent MCMC. The aim of this project is to provide a flexible structure for the fast simulation of Bayesian adaptive designs. We focus our attention on multi-arm multi-stage (MAMS) designs investigated as a first step. We will illustrate how the BATS package (Bayesian Adaptive Trials Simulator) can be used to define the operating characteristics of a Bayesian adaptive design for different types of endpoints given the most common adaptations - stopping trial/arms for efficacy or futility, fixed or (covariate-adjusted) response-adaptive randomisation - based on self-defined rules. Other important features include: parallel processing, customisability, use on a cluster computer or PC/Mac, adjustment for covariates.

BATS has been successfully used for the recent rounds of MRFF or NHMRC grant applications.

Tuesday, 05/Sept/2023 2:20pm - 2:40pm
ID: 248 / S31: 2
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Interpretable machine learning in biostatistics: Methods, applications and perspectives
Keywords: Interpretable Machine Learning, Model-agnostic, Counterfactual Explanations, Multi-Objective Optimization, R package

Multi-Objective Counterfactual Explanations

Susanne Dandl[1,2], Andreas Hofheinz[1], Martin Binder[1,2], Bernd Bischl[1,2], Giuseppe Casalicchio[1,2]
[1]LMU Munich; [2]Munich Center for Machine Learning (MCML); susanne.dandl@stat.uni-muenchen.de

In recent years, various methods have been proposed to make complex prediction models explainable. A method that explains the prediction of a point of interest in the form of "what if" statements are counterfactual explanations, or short counterfactuals. In the medical context, counterfactuals allow for statements such as "If you do not have diabetes and have a BMI of 25 instead of 30, your model-predicted risk of chronic heart disease would drop from 75% to 40%".

To be a counterfactual, a generated point must have (some of) the following properties: (1) The point's prediction should be equal to the desired prediction, (2) it should be close to the point of interest, (3) only a few feature changes should be proposed, and (4) the point should adhere to the data manifold.

Based on these desired properties, an optimization problem can be formulated to generate counterfactuals. In our work, we argue that this optimization task is inherently multi-objective. This is because some of the properties contradict each other (e.g., in order to reach a desired prediction, more feature changes are necessary), while all properties are equally important. Therefore, rather than just one counterfactual, a whole set of equally good counterfactuals for a point of interest exists.

With this in mind, we developed the multi-objective counterfactuals (MOC) method, which is model-agnostic and works for all types of features. On a dataset on indicators for chronic heart disease (Centers for Disease Control and Prevention), we demonstrate how the method provides interesting insights into the underlying predictive model. Therefore, we use the MOC implementation in the counterfactuals R package available on CRAN.

Since a large set of returned counterfactuals can be overwhelming for users, we also address how numerous counterfactuals can be visualized, what options users have to select individual counterfactuals, and how multiple similar counterfactuals can be combined.

ID: 458 / **S57: 4**
Presentation Submissions - Invited Session
Invited Sessions: Causal inference and the art of asking meaningful questions
Keywords: mediation analysis, heterogeneity, estimands

On the choice of estimands when the role of an intermediate variable is of interest

Rhian Mair Daniel
Cardiff University, United Kingdom; danielr8@cardiff.ac.uk

In this talk, I will discuss different aspects of estimand choice in the presence of an intermediate (or mediating) variable of scientific interest. This will include an overview of the subtle differences between different flavours of direct and indirect effects already suggested in the mediation analysis literature. I will also discuss new perspectives on the role of heterogeneity, and its consequences for the transportability of mediation estimands (as well as total effect estimands) across different populations, in particular when there are causal effects in qualitatively opposite directions along different pathways from exposure to outcome.

Tuesday, 05/Sept/2023 3:00pm - 3:20pm
ID: 326 / **S29: 4**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Generalized pairwise comparisons

Generalized pairwise comparisons as a pragmatic alternative to non-inferiority trial designs

Mickaël De Backer, Samuel Salvaggio, Vaiva Deltuvaite-Thomas, Sarah Kosta, Emilie Barré, Jean-Christophe Chiem, Everardo Saad, Marc Buyse
International Drug Development Institute, Belgium; mickael.de.backer@iddi.com

In many clinical situations, the medical question of interest requires the conduct of a non-inferiority trial (NI), but the latter are often unfeasible in addition to bringing several challenges in contrast to superiority trials. In this presentation, we examine the use of generalized pairwise comparisons (GPC) as a pragmatic alternative to NI trials for addressing the problem of ensuring efficacy and tolerability. The method of GPC is a recent proposal that allows the simultaneous evaluation of several outcomes of interest that can be of any type. These outcomes can further be prioritized to reflect one's opinion regarding their perceived hierarchy of clinical importance. As an illustration, we consider the design of a randomized trial for patients with acute promyelocytic leukemia, where reducing the standard treatment dose is intended to improve tolerability. We highlight the different steps and choices for designing a trial constructed on GPC, based here in particular on historical data. The latter form the basis of a thorough simulation exercise for sample size determination. This presentation thus highlights how GPC can be considered an essential tool for assessing efficacy and tolerability in scenarios where NI trials are difficult to conduct, particularly when researching vulnerable groups.

Tuesday, 05/Sept/2023 12:00pm - 12:20pm
ID: 228 / **S22: 3**
Presentation Submissions - Invited Session
Invited Sessions: Net benefit, win odds, and win ratio: Methods, analysis, and interpretation

Inferential methods for generalized pairwise comparisons of censored data

Vaiva Deltuvaite-Thomas
IDDI, Belgium; vaiva.thomas@gmail.com

The family of Generalized Pairwise Comparisons (GPC) include statistics/methods that are generalizations of Wilcoxon-Mann-Whitney test. Multiple extensions of Wilcoxon-Mann-Whitney test of and other GPC methods have been proposed to handle censored data. These methods differ in handling the loss of information due to censoring: ignoring non-informative pairwise comparisons (Gehan, Harrell and Buyse); imputing predicted scores using estimates of the survival distribution (Efron, Péron and Latta); or inverse probability of censoring weighting (IPCW, Datta and Dong). In our talk, we will present available GPC methods for censored data, discuss how each of them influences the operational characteristics of the GPC tests, and provide recommendations related to their choice in various situations.

Monday, 04/Sept/2023 2:20pm - 2:40pm
ID: 344 / S10: 2
Presentation Submissions - Regular Session (Default)
Topic of Submission: Biomarkers and diagnostics, Statistical modelling (regression modelling, prediction models, …), Personalized health care, Machine Learning and Data Science
Keywords: Acute myocardial infarction, biomarker, validation, machine learning, probability, super learner

Personalized diagnosis in suspected myocardial infarction: the ARTEMIS study

Eleonora Di Carluccio[9], Johannes Tobias Neumann. MD[1,2,3,4], Francisco Ojeda. PhD[1,3], Raphael Twerenbold. MD[1,2,3,5], Andreas Ziegler. PhD[1,9,29], Sally J. Aldous. MD[6], Brandon R. Allen. MD[7], Fred S. Apple. PhD[8], Hugo Babel. PhD[9], Robert H. Christenson. MD[10], Louise Cullen. MD[11], Dimitrios Doudesis. PhD[12], Ulf Ekelund. MD. PhD[13], Evangelos Giannitsis. MD[14], Jaimi Greenslade. PhD[11], Kenji Inoue. MD[15], Tomas Jernberg. MD[16], Peter Kavsak. PhD[17], Till Keller. MD[18], Kuan Ken Lee. MD[12], Bertil Lindahl. MD[19], Thiess Lorenz[1,2,3], Simon A. Mahler. MD[20], Nicholas L. Mills. MD[12], Arash Mokhtari. MD[21], William Parsonage. DM[22], John W. Pickering. PhD[23], Christopher J. Pemberton. PhD[24], Christoph Reich. MD[25], A. Mark Richards. MD[23], Yader Sandoval. MD[26], Martin P. Than. MD[27], Betül Toprak. MD[1,2,3,5], Richard W. Troughton. MD[24], Andrew Worster. MD[28], Tanja Zeller. PhD[1,2,3,5], Stefan Blankenberg. MD[1,2,3]

[1]Department of Cardiology, University Heart and Vascular Center, University Medical Center Hamburg-Eppendorf, Hamburg, Germany; [2]German Center for Cardiovascular Research (DZHK), Partner Site Hamburg/Kiel/Lübeck, Hamburg, Germany; [3]Population Health Research Department, University Heart and Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany; [4]Department of Epidemiology and Preventive Medicine, School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia; [5]University Center of Cardiovascular Science, University Heart and Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany; [6]Department of Cardiology, Christchurch Hospital, New Zealand; [7]Department of Emergency Medicine, College of Medicine, University of Florida, Gainesville, FL, USA; [7]Department of Emergency Medicine, College of Medicine, University of Florida, Gainesville, FL, USA; [8]Departments of Laboratory Medicine and Pathology, Hennepin Healthcare/HCMC and University of Minnesota, Minneapolis, MN, USA; [9]Cardio-CARE, Medizincampus Davos, Davos, Switzerland; [10]Department of Pathology, University of Maryland School of Medicine, Baltimore, MD, USA; [11]Department of Emergency Medicine, Royal Brisbane and Women's Hospital, Herston, Queensland, Australia; [12]BHF Centre for Cardiovascular Science, University of Edinburgh, Edinburgh, United Kingdom; [13]Lund University, Skåne University Hospital, Department of Internal and Emergency Medicine, Lund, Sweden; [14]Department of Cardiology, Heidelberg University Hospital, Heidelberg, Germany; [15]Juntendo University Nerima Hospital, Tokyo, Japan; [16]Department of Clinical Sciences, Danderyd University Hospital, Karolinska Institutet, Stockholm, Sweden; [17]Department of Pathology and Molecular Medicine, McMaster University, Hamilton, ON, Canada; [18]Department of Cardiology, Kerckhoff Heart and Thorax Center, Bad Nauheim, Germany; [19]Department of Medical Sciences and Uppsala Clinical Research Center, Uppsala University, Sweden; [20]Department of Emergency Medicine, Wake Forest School of Medicine, Winston-Salem, NC, USA; [21]Department of Internal Medicine and Emergency Medicine and Department of Cardiology, Lund University, Skåne University Hospital, Lund, Sweden; [22]Australian Centre for Health Service Innovation, Queensland University of Technology, Kelvin Grove, Australia; [23]Department of Medicine, University of Otago Christchurch and Emergency Department, Christchurch Hospital, Christchurch, New Zealand; [24]Department of Medicine, Christchurch Heart Institute, University of Otago, New Zealand; [25]Department of Cardiology, Heidelberg University Hospital, Heidelberg, Germany; [26]Minneapolis Heart Institute, Abbott Northwestern Hospital, and Minneapolis Heart Institute Foundation, Minneapolis, MN, USA; [27]Department of Medicine, University of Otago Christchurch and Emergency Department, Christchurch Hospital, Christchurch, New Zealand; [28]Division of Emergency Medicine, McMaster University, Hamilton, ON, Canada; [29]School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Pietermaritzburg, South Africa.;
Eleonora.diCarluccio@cardio-care.ch

In suspected myocardial infarction (MI), guidelines recommend using high-sensitivity cardiac troponin (hs-cTn)-based approaches. These require fixed assay-specific thresholds and timepoints, without directly integrating clinical information. Using machine-learning techniques including hs-cTn and clinical routine variables, we developed and validated a model to estimate the individual probability of MI, while allowing for numerous hs-cTn assays. The aim of this presentation is to describe the approach for developing and validating this diagnostic model.

In 2,575 patients presenting to the emergency department with suspected MI, two ensembles of machine-learning models using single or serial concentrations of six different hs-cTn assays were derived to estimate the individual MI probability. Twelve routinely available variables including age, sex, cardiovascular risk factors, electrocardiography, and hs-cTn were included in the ARTEMIS models. First, multiple imputation was performed. Second, full models were trained using ten-fold cross-validation. Third, variable selection was performed by using the information from all hs-cTn models. Fourth, reduced models were trained using ten-fold cross-validation. Fifth, a superlearner with equal weights was estimated. Model performance was assessed using the logLoss. It was validated in an external cohort with 1,688 patients and tested for global generalizability in 13 international cohorts with 23,411 patients after calibration.

Model performance of the reduced models was superior to the full model and superior to the hs-cTn only-model in the training data. It performed best on the validation and generalization data, and it was significantly better than the hs-cTn only-model. Models based on the superlearner generally outperformed the single learners. Using a single hs-cTn measurement, the ARTEMIS model allowed direct rule-out of MI with very high and similar safety but up to tripled efficiency compared to the guideline-recommended strategy.

We developed and validated diagnostic models to accurately estimate the individual probability of MI, which allow for variable hs-cTn use and flexible timing of resampling. The development of models using different hs-cTn assays led to substantial greater stability in the model performance due to improved variable selection properties. Furthermore, the use of an equal-weights superlearner further increased the stability of the machine learning models.

This is a joint contribution by the ARTEMIS team which all authors are part of.

Wednesday, 06/Sept/2023 9:10am - 9:30am
ID: 177 / **S44: 2**
Presentation Submissions - Invited Session
Invited Sessions: Online hypothesis testing and subgroup analyses in complex innovative designs
Keywords: biomarker detection, conditional p-values, data-adaptive multiple test, false discovery rate, stochastic orders

Multiple testing of partial conjunction null hypotheses, with application to replicability analysis of high-dimensional studies

Thorsten Dickhaus[1], **Ruth Heller**[2], **Anh-Tuan Hoang**[1], **Anna Vesely**[1]
[1]University of Bremen, Germany; [2]Tel Aviv University; dickhaus@uni-bremen.de

The partial conjunction null hypothesis is tested in order to discover a signal that is present in multiple (sub-)studies. The standard approach of carrying out a multiple test procedure on the partial conjunction *p*-values can be extremely conservative. We suggest alleviating this conservativeness, by eliminating many of the conservative partial conjunction *p*-values prior to the application of a multiple test procedure. This leads to the following two-step procedure: first, select the set with partial conjunction *p*-values below a selection threshold; second, within the selected set only, apply a family-wise error rate or false discovery rate controlling procedure on the conditional partial conjunction *p*-values, where conditioning is on the selection event. We discuss theoretical properties of the proposed procedure, and we demonstrate its performance on simulated and real data.

Monday, 04/Sept/2023 11:25am - 11:50am
ID: 457 / **S4: 2**
Presentation Submissions - Featured Session
Featured Sessions: Biometrical Journal Showcase - Editor's Selection
Keywords: hazard ratio order, meta-analysis, proportion of true null hypotheses, Schweder–Spjøtvoll estimator

Randomized p-values in replicability analysis

Thorsten Dickhaus, **Anh-Tuan Hoang**
University of Bremen, Germany; dickhaus@uni-bremen.de

We will be concerned with testing replicability hypotheses for many endpoints simultaneously. This constitutes a multiple test problem with composite null hypotheses. Traditional p-values, which are computed under least favourable parameter configurations (LFCs), are over-conservative in the case of composite null hypotheses. As demonstrated in prior work, this poses severe challenges in the multiple testing context, especially when one goal of the statistical analysis is to estimate the proportion π_0 of true null hypotheses. To address this issue, we will discuss the application of randomized p-values in replicability analysis. By means of theoretical considerations as well as computer simulations, we will demonstrate that their usage typically leads to a much more accurate estimation of π_0 than the LFC-based approach. Furthermore, we will draw connections to other recently proposed methods for dealing with conservative p-values in the multiple testing context. Finally, we will present a real data example from genomics.

Monday, 04/Sept/2023 4:10pm - 4:30pm
ID: 133 / **S18: 1**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Free Contributions
Keywords: patient preference study, discrete choice experiment, swing weighting, case studies.

How patient preference studies can support decision-making in early drug development (Phases 1-2)

Sheila Dickinson, Byron Jones, Nigel Cook
Novartis, Switzerland; sheila.dickinson@novartis.com

Two case studies will be presented to demonstrate how patient preference studies can support decision-making in early drug development (Phases 1-2). The case studies used two different quantitative types of preference study: a Discrete Choice Experiment and a Swing Weighting (SW) approach.

The first case study used a DCE and obtained preferences from patients suffering from Chronic Obstructive Pulmonary Disease (COPD) [1]. It involved a total of 1050 patients from five countries and aimed to understand the relative importance to patients of different endpoints and to inform the choice of clinical, PRO, and digital endpoints for inclusion in pivotal clinical trials in COPD.

The second case study [2], which used the SW approach, is a patient preference study in a rare chronic kidney disease, IgA Nephropathy. Here the interview-based SW methodology was used to enable robust preference data to be derived from a small number of patients. The results of the study were used to evaluate the relative importance to patients of different benefits and risks, including the trade-offs that patients are willing to make.

References

1. Cook, N.S., Criner, G.J., Burgel, P-R, Mycock, M., Gardener, T., Mellor, P. Hallworth, P., Sully, K., Tatlock, S., Klein, B., Jones, B., Le Rouzic, O., Adams, K., Phillips, K., McKevitt, M. Toyama, K. and Gutzwiller, F. (2022). People living with moderate-to-severe COPD prefer improvement of daily symptoms over the improvement of exacerbations: a multicountry patient preference study. ERJ Open Research, 8(2):686-2021. DOI: 10.1183/23120541.00686-2021
2. Marsh, K., Ho, K-A., Lo, R., Zaour, N., George, A.T. and Cook, N.S. (2021). Assessing Patient Preferences in Rare Diseases: Direct Preference Elicitation in the Rare Chronic Kidney Disease, Immunoglobulin A Nephropathy. The Patient, 14(6), 837-847. doi: 10.1007/s40271-021-00521-3.

Monday, 04/Sept/2023 2:00pm - 2:40pm
ID: 205 / **S9: 1**
Presentation Submissions - Invited Session
Invited Sessions: A causal inference perspective on estimands in clinical trials
Keywords: causal inference, estimands, real-world data, target trial emulation

From causal inference with observational data to estimands in RCTs and back!

Vanessa Didelez
Leibniz Institute for Prevention Research and Epidemiology - BIPS, Germany; vdidelez@uni-bremen.de

In this presentation, I will start with an overview on common estimands used in causal analyses of observational (aka 'real world') data, their motivation, and the key structural assumptions underlying typical analytical methods. Applied examples range from estimating the (side)effects of therapies or the effectiveness of cancer screening using electronic health data, to evaluating lifestyle recommendation aimed at preventing childhood obesity. In most of these examples, the relevant exposure or treatment is not a binary point variable but a sustained or time-varying exposure/treatment, and the outcome is longitudinal or a time-to-event. This means the estimand should be carefully formulated in view of what sustained, time-varying or possibly adaptive treatment strategies we wish to compare. While the key assumptions, required for most methods, of sufficient measured time-varying confounder information and sufficient overlap between observed and targeted strategies often appear hard to defend, approaches to strengthen their plausibility have been developed and are increasingly applied.

Secondly, I will discuss possible implications and lessons-to-be-learned from observational causal inference for the design and analyses of RCTs with typical intercurrent events as addressed by the ICH E9 Addendum. The latter can often be seen to alter either the intended meaning of treatment or of the outcome, so that relevant estimands must be formulated. This will often take us back to dynamic or adaptive treatment strategies. However, sometimes, notions of direct effects are invoked, and I will discuss why these are more problematic to interpret. In this context, and with view to drug development, I will explain how so-called separable effects might be an interesting alternative. Particular attention will be paid to how different estimands, in observational studies or RCTs, inform different decision makers, such as individuals, physicians or public health authorities. A further key difference, compared to most sources of observational data, is that in RCTs (and to some extent in studies using real-world data) we have more influence on what information will be available for the analysis, and thus can ensure that a rich set of confounders is measured as well as other useful information enabling sensitivity analyses.

Finally, I will come full circle and address the role of formulating a target trial, i.e. a somewhat idealized hypothetical trial, in the analysis of observational / real-world data. Target trial emulation is an important principle especially for eliciting actionable estimands in complex longitudinal data situations, but also for avoiding self-inflicted biases in the analysis of observational data. Thus, on the one hand, the analysis of RCTs with intercurrent events looks to learn from approaches to causal inference developed for observational data; on the other hand, the causal analysis of observational data can be much strengthened by adhering to certain design principles of randomized trials.

The presentation will focus on general principles, examples and interpretation more than on technical details.

Monday, 04/Sept/2023 11:00am - 11:25am
ID: 454 / **S4: 1**
Presentation Submissions - Featured Session
Featured Sessions: Biometrical Journal Showcase - Editor's Selection
Keywords: estimands; confounding; odds ratio; hazard ratio

On the logic of collapsibility for causal effect measures

Vanessa Didelez[1], **Mats Stensrud**[2]
[1]Leibniz Institute for Prevention Research and Epidemiology - BIPS, Germany; [2]EPFL; vdidelez@uni-bremen.de

There is a long history of confusing "non-collapsibility" and "confounding", including a long history of attempts to clarify the distinction. The topic has received renewed attention in the context of subgroup analyses in randomized trials together with the issue of choosing an estimand in view of intercurrent events. The problem is compounded by the fact that the typical examples of non-collapsible measures, odds-ratios and hazard-ratios, also have a problematic causal interpretation, which is again a separate issue from whether they are affected by confounding in a given study. We discuss these issues from a causal point of view, separating it from whether the basis of inference is trial- or real-world data.

The key messages are: (1) To avoid misunderstandings, associational concepts of dependence should clearly and formally be distinguished from causal contrasts. (2) Confounding and non-collapsibility are separate issues and should be kept apart, and an RCT does guarantee non-confounding at baseline (but can suffer from many other problems). (3) Odds ratios and especially hazard ratios are problematic as causal contrasts, the latter due to inherent conditioning on survival, and also e.g. for transportability. However, collapsibility does not in itself guarantee a meaningful causal contrast, e.g. hazard differences. Moreover, there is empirical evidence that patients and domain experts prefer causal contrasts in terms of absolute risk, and their advantages and computation will be illustrated in this talk.

Thursday, 07/Sept/2023 10:40am - 11:00am
ID: 160 / **S69: 1**
Presentation Submissions - Regular Session (Default)
Topic of Submission: High dimensional data, genetic and x-omics data
Keywords: convex optimization, dynamic networks, graphical models, high-dimensionality, TCGA

High-dimensional graphical models varying with multiple external covariates

Louis Dijkstra, Ronja Foraita
Leibniz Institute for Prevention Research and Epidemiology - BIPS, Bremen, Germany; dijkstra@leibniz-bips.de

High-dimensional networks play a key role in understanding complex relationships. These relationships are often dynamic in nature and can change with multiple external factors (e.g., time and case-control status). Methods for estimating graphical models are often restricted to static graphs or graphs that can change with a *single* covariate (e.g., time). We propose a novel class of graphical models, the covariate-varying network (CVN), that can change with *multiple* external covariates. This extension sounds trivial at first; however, it poses serious conceptual and computational challenges.

In order to introduce sparsity, we apply a L_1 penalty on the precision matrices of $m \geq 2$ graphs we want to estimate. These graphs often show a level of similarity (i.e., the graphs are 'smooth'). This smoothness is modelled using a 'meta-graph' with m nodes, each corresponding to a graph one wants to estimate. The (weighted) adjacency matrix of the meta-graph represents the strength with which similarity is enforced between the m graphs.

The resulting optimization problem is solved by employing an alternating direction method of multipliers (ADMM). One update-step in the resulting ADMM requires one to repeatedly solve a 'weighted fused signal approximator', which, to the best of our knowledge, had not been solved before. We do this by reformulating it as a Generalized LASSO problem and solving it with an ADMM developed specifically for this task.

We test our method using a simulation study and we show the method's applicability by analyzing the dependence structure of gene expressions within the p53 pathway of head and neck squamous cell carcinoma patients; a dataset from The Cancer Genome Atlas (TCGA; https://www.cancer.gov/tcga) with tumor stage and tumor site as external covariates.

Tuesday, 05/Sept/2023 12:20pm - 12:40pm
ID: 125 / **S26: 5**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical hypothesis testing (covariate adjustment, nonparametric methods, multiple comparisons, …)
Keywords: Online multiple testing, Discrete hypothesis testing, Weighted hypothesis testing, False discovery rate

Online multiple testing with heterogeneous data

Sebastian Doehler[1], **Iqraa Meah**[1,2], **Etienne Roquain**[2]
[1]Darmstadt University of Applied Sciences, Germany; [2]Universite Sorbonne, France; sebastian.doehler@h-da.de

Online multiple testing refers to the setting where a possibly infinite number of hypotheses are tested, and the p-values are available one by one sequentially. This differs from classical multiple testing where the number of tested hypotheses is finite and known beforehand, and the p-values are available simultaneously.

It is well-known that the existing methods for online multiple testing can suffer from a significant loss of power if the null p-values are conservative. In this work, we extend the previously introduced methodology to obtain more powerful procedures for the case of super-uniformly distributed p-values. These types of p-values arise in important settings, e.g. when discrete hypothesis tests are performed or when the p-values are weighted. To this end, we introduce the method of superuniformity reward (SUR) that incorporates information about the individual null cumulative distribution functions. Our approach yields several new 'rewarded' procedures that offer uniform power improvements over known procedures and come with mathematical guarantees for controlling online error criteria based either on the family-wise error rate (FWER) or the marginal false discovery rate (mFDR).

Tuesday, 05/Sept/2023 11:00am - 11:20am
ID: 132 / **S26: 1**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Time-to-Event Analysis
Keywords: survival analysis, multiple contrast tests, non-proportional hazards, weighted log-rank test

multiCASANOVA - Multiple group comparisons for non-proportional hazard settings

Ina Dormuth[1], Frank Konietschke[2], Carolin Herrmann[2], Markus Pauly[1], Marc Ditzhaus[3]
[1]TU Dortmund University, Germany; [2]Charité – Universitätsmedizin Berlin, Germany; [3]Otto von Guericke University Magdeburg, Germany; ina.dormuth@tu-dortmund.de

Comparing multiple groups based on time-to-event data is a common subject of interest in clinical studies. The log-rank test is commonly used to assess differences between groups (e.g., treatment groups) and is optimal under the assumption of proportional hazards. However, when this assumption is violated, the log-rank test dramatically loses power. When comparing two groups, various methods that are more robust towards a violation of the proportional hazards assumption have already been proposed. One promising approach is the combination of several weighted log-rank tests, such as CASANOVA [1]. Nevertheless, when multiple groups are compared to one another, it is often not only of interest to discover a global significant difference but to determine the origin of this difference. This results in the necessity to test multiple individual hypotheses. In order to obtain a test decision for the individual comparisons, the log-rank test requires adjustment for multiple testing, such as Bonferroni. Such corrections control the familywise type I error but are usually conservative. We propose a new multiple contrast test based on the CASANOVA approach [1]. This makes the proposed test more powerful under non-proportional hazards and at the same time, renders the need for p-value correction obsolete. We evaluate the performance of the test in extensive Monte-Carlo Simulation studies covering proportional as well as non-proportional hazards settings.

Reference:

[1] Ditzhaus, M., Genuneit, J., Janssen, A., & Pauly, M. (2021). CASANOVA: Permutation inference in factorial survival designs. *Biometrics*.

Monday, 04/Sept/2023 3:20pm - 3:40pm
ID: 219 / **S13: 5**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), Preclinical drug development, safety and toxicology
Keywords: Functional Shrinkage, Subspace shrinkage, Bayes, Non-parametric, Dose-Response

Bayesian nonlinear functional subspace shrinkage with application to gene expression dose-response data

Julia Christin Duda[1], **Matthew Wheeler**[2]
[1]TU Dortmund University, Germany; [2]National Institute of Environmental Health Sciences, United States; duda@statistik.tu-dortmund.de

Background: Shrinkage estimation has become very successful over the last decades with ridge or lasso estimators being the most prominent examples. With its natural Bayesian flavor, a large literature focuses on Bayesian shrinkage methods. Recently, Shin et al. (2020) extended parameter shrinkage to (linear) functional subspace shrinkage in the regression context by presenting the functional horseshoe (fHS) prior. Using flexible semiparametric models such as splines, the fHS prior induces shrinkage of the entire shape of the regression function towards a class of parametric models. Wiemann and Kneib (2021) moved forward this idea by incorporating the aspect of smoothness into the prior as an additive component.

Method: We propose a functional shrinkage method that shrinks into nonlinear subspaces through linear approximation while smoothness is added as a subspace but not as an additive penalty matrix in the prior. This approach allows deviations from the subspace in a weighted manner if the data locally calls for it.

We demonstrate our approach on gene expression dose-response data where human embryonic stem cells are exposed to different concentrations of valproic acid. Gene expression dose-response data is known to deviate from well-established parametric curves, but this is typically only in localized regions. As nonlinear subspace, we therefore select the mechanistically motivated four-parameter log-logistic model, also known as Hill or sigmoidal Emax model. We compare our method to parametric smoothing splines and Bayesian P-splines in a simulation study.

Results: In the application, the proposed method maintains its desired characteristics. It shrinks the dose response curves into the biologically plausible model, while it locally allows reasonable, data-driven deviations. This makes our method competitive with the other approaches in the simulation study.

References:

1. Shin, M., Bhattacharya, A., & Johnson, V. E. (2020). Functional horseshoe priors for subspace shrinkage. *Journal of the American Statistical Association*, *115*(532), 1784-1797.
2. Wiemann, P., & Kneib, T. (2021). Adaptive shrinkage of smooth functional effects towards a predefined functional subspace. *arXiv preprint arXiv:2101.05630*.

Thursday, 07/Sept/2023 9:30am - 09:50am
ID: 443 / S58: 5
Presentation Submissions - Regular Session (Default)
Topic of Submission: Clinical trial designs (adaptive and platform designs, external controls, …)
Keywords: Innovative clinical trial design, RCT augmentation with close-to-real-world elements, Health Technology Assessments

Generating the right evidence at the right time: Principles of a new class of flexible augmented clinical trial designs

Cornelia Dunger-Baldauf[1], Rob Hemmings[3], Frank Bretz[1], Byron Jones[2], Anja Schiel[5], Chris Holmes[4]
[1]Novartis Pharma AG, Switzerland; [2]Novartis Pharma AG, UK; [3]Consilium Salmonson&Hemmings; [4]Oxford University; [5]Norwegian Medicines Agency; cornelia.dunger-baldauf@novartis.com

To support informed decision making by pharmaceutical companies, regulators, health technology assessment (HTA) bodies, payers, patients and physicians, clear descriptions of the benefits and risks of a treatment for a given medical condition should be made available in a timely fashion. Historically, pharmaceutical drug development proceeded in a sequential fashion, focusing foremost on authorisation of a new treatment prior to addressing questions relevant to other stakeholders involved in getting new medicines to patients. However, data generated later, perhaps through observational studies, can be difficult to compare with earlier randomised trial data, resulting in confusion in understanding and interpretation of treatment effects. Moreover, the scientific questions these later experiments can serve to answer often remain vague.

We propose FACTIVE (Flexible Augmented Clinical Trial for Improved eVidence gEneration), a new class of study designs enabling flexible augmentation of confirmatory randomised controlled trials with concurrent and close-to -real-world elements. Our starting point is to use clearly defined objectives for evidence generation, which are formulated through early discussion with HTA bodies and are additional to regulatory requirements for authorisation of a new treatment. These enabling designs facilitate estimation of certain, well-defined treatment effects in the confirmatory part and other, complementary treatment effects in a concurrent real-world part. Each stakeholder should use the evidence that is relevant within their own decision-making framework. High quality data are generated under one single protocol and the use of randomisation ensures rigorous statistical inference and interpretation within and between the different parts of the experiment.

Wednesday, 06/Sept/2023 10:40am - 11:00am
ID: 187 / S54: 1
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …)
Keywords: MMRM, PROC MIXED, R, glmmTMB, nlme, mmrm

Comparing R libraries with SAS's PROC MIXED for the analysis of longitudinal continuous endpoints using MMRM

Gonzalo Duran-Pacheco[1], Julia Dedic[2], Philippe Boileau[3]
[1]Roche, Switzerland; [2]Roche, Canada; [3]Genentech, US; gonzalo_christian.duran_pacheco@roche.com

Mixed-effect models for repeated measures (MMRM) are widely accepted as the statistical approach to analyse longitudinal continuous endpoints in clinical trials. The clinical outcome is observed in study patients over time, making values within subjects more similar than values of other individuals, which implies within-subject correlation. MMRM accounts for such correlation by modeling explicitly the corresponding covariance structure which can be specified in alternative ways (unstructured, compound symmetry, first order autoregressive, Toeplitz, antedependence, and others). SAS's PROC MIXED has been a broadly accepted software to implement MMRM in clinical trials. However, due to growing popularity in academia, industry and increased compliance with regulations and requirements, the open source software R has become incrementally favored in clinical trial settings. In this study we compare MMRM results obtained by three R libraries: nlme, glmmTMB and mmrm, versus SAS PROC MIXED in an attempt to reproduce clinical study report results of five phase-3 clinical trials and we also conduct simulations. We will report results regarding execution time, marginal expected means, contrasts, corresponding standard errors and p-values. We will also report results of MMRM using various alternative covariance structures.

Wednesday, 06/Sept/2023 9:30am - 9:50am
ID: 336 / **S49: 4**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Epidemiology, Time-to-Event Analysis, Statistical hypothesis testing (covariate adjustment, nonparametric methods, multiple comparisons, …)
Keywords: pharmacology, signal detection, time-to-event models, Bayesian test, generalized Weibull distribution

Signal detection of adverse drug reactions: The Bayesian power generalized Weibull shape parameter test

Julia Alexandra Dyck
Bielefeld University, Germany; j.dyck@uni-bielefeld.de

After the release of a drug on the market, pharmacovigilance monitors the occurrence and changes in known adverse drug reactions (ADRs) as well as detects new ADRs in the population. This is done to keep a drug's harm profile updated and can potentially result in adjustments of the prescription labeling or – in the extreme case – a recall of the product from the market. In recent years the interest in the use of longitudinal electronic health records for pharmacovigilance increased. Cornelius et. al (2012) provided a signal detection test based on the Weibull distribution shape parameter. Sauzet and Cornelius (2022) refined this approach, proposing a test based on the power generalized Weibull distribution shape parameters (PgWSP). The power generalized Weibull (PgW) distribution is characterized by a scale parameter and two shape parameters. If both shape parameters of the PgW distribution are equal to one, the distribution reduces to an exponential distribution with constant hazard over time. A constant hazard is interpreted as no temporal association between a drug and an adverse event.

Signal detection can be improved by incorporating existing knowledge about the ADR profile of drugs from the same family or based on expert knowledge about the drug mechanism. Therefore, we propose the development of a Bayesian PgWSP test. The hypothesis test in the Bayesian context is based on a region of practical equivalence (ROPE) reflecting the shape parameters' values under the null hypothesis (Krushke, 2015). For each parameter, a credibility interval deducted from the posterior density is compared to the ROPE. If the intersection between ROPE and credibility interval is empty for at least one shape parameter, a signal is raised.

We performed a simulation study with the aim to find the optimal ROPE and credibility interval for signal detection using a Bayesian PgWSP approach. Samples are generated with varying sample sizes, background rates reflecting the number of symptom observations in the general population, and proportions of adverse event observations caused by a drug. For the scale parameter, we test a fixed prior value, a gamma, and a lognormal prior distribution. For both shape parameters, either gamma or lognormal prior distributions are used. The priors are characterized in terms of mean and standard deviation based on hypothetical prior assumptions. Prior assumptions considered are: the symptom is no ADR of the drug, the symptom is an ADR of the drug with the highest risk of occurrence either at the beginning, in the middle, or at the end of the observation period. ROPEs of various types or widths are considered. For posterior credibility intervals, we use either equal-tailed or highest-density intervals of the posterior distributions with varying credibility levels. The optimal tuning parameters are determined based on the area under the curve.

Monday, 04/Sept/2023 4:50pm - 5:10pm
ID: 198 / S18: 3
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical hypothesis testing (covariate adjustment, nonparametric methods, multiple comparisons, …)
Keywords: Assurance, Prior Distribution, Three-way Similarity Testing, Multiple Coprimary Endpoints

Assurance of three-way PK (PD) Biosimilarity studies with multiple coprimary endpoints

Rachid El Galta[1], Elise Burmeister-Getz[3], Jessie Wang[2], Susanne Schmitt[1], Ramin Arani[2], Arne Ring[1]
[1]HEXAL AG, Germany; [2]Sandoz Pharmaceutical, USA; [3]Novartis Institutes for BioMedical, USA; rachid.el_galta@sandoz.com

In Biosimilar development, a human PK (PD) clinical pharmacology study is an essential step in the stepwise approach for demonstrating biosimilarity. Such a study is commonly designed to demonstrate 3-way PK (PD) similarity between a Biosimilar candidate, a US approved reference and an EU approved reference in multiple coprimary endpoints (e.g., AUCinf, AUClast, Cmax) with high confidence. For each of the coprimary endpoints the null hypothesis of no-similarity is tested based on Two One-Sided test (TOST) for all 3 pairwise comparisons with respect to the similarity margins. In the calculation of the required study sample size, unknown expected treatment differences and covariance parameters of the coprimary endpoints are usually substituted by point estimates (guestimates) from historical data whenever available, while ignoring their uncertainty. However, misspecification of these parameters is likely to result in an underpowered study and hence a lower actual trial probability of success.

For bioequivalence studies with one primary endpoint, Ring et al. (2019) proposed the use of the assurance approach to account for uncertainty on unknown treatment difference, especially using the expectation of the power function with respect to the prior distribution of the treatment difference.

In this presentation we show how to extend the assurance approach for multivariate coprimary endpoints for testing three-way similarity by estimating the posterior distribution of the study power as a function of the vector of the treatment differences and the covariance matrix using multivariate normal and Wishard prior distributions, respectively. In addition, we discuss how to inform the prior distributions of the unknow parameters for the following scenarios: 1) Historical PK (PD) data available for all three arms; 2) Historical available in one reference arm; and 3) Historical is not available. We also implemented the approach using an R-shiny application.

Tuesday, 05/Sept/2023 12:00pm - 12:20pm
ID: 340 / **S23: 4**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), Real world data and evidence, Time-to-Event Analysis

A general estimation framework for multistate survival processes with flexible specification of the transition intensities

Alessia Eletti[1], **Giampiero Marra**[1], **Rosalba Radice**[2]
[1]University College London, United Kingdom; [2]Bayes Business School, United Kingdom; alessia.eletti.19@ucl.ac.uk

When interest lies in the progression of a disease rather than on a single outcome, multistate Markov models represent a natural and powerful modelling approach. Often the nature of the phenomenon itself renders constant monitoring unfeasible, thus leading to the process being observed only intermittently. This setting is challenging and existing methods and their implementations do not yet provide flexible enough mechanisms for fully exploiting the information contained in the data. To this end, we propose a closed-form expression for the local curvature information of the transition probability matrix, which has not been previously attempted. Building on this, we introduce a general framework that allows one to model any type of multistate process where the transition intensities are flexibly specified as functions of regression splines. Parameter estimation is carried out through a carefully structured, stable penalised likelihood approach. We exemplify our method by modelling cognitive decline in the English Longitudinal Study of Ageing. To support applicability and reproducibility, all developed tools are implemented in the R package flexmsm.

Wednesday, 06/Sept/2023 9:50am - 10:10am
ID: 390 / S48: 5
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), Statistical hypothesis testing (covariate adjustment, nonparametric methods, multiple comparisons, …), Preclinical drug development, safety and toxicology
Keywords: power analysis, sample size planning, study design, uncertainty, analytical flexibility

Researcher Degrees of Freedom in Power Analyses and Sample Size Planning

Nicole Ellenbach[1], Anne-Laure Boulesteix[1], Sabine Hoffmann[2], Bruno L. Cadilha[3], Sebastian Kobold[3,4,5], Juliane C. Wilcke[1]

[1]Institute for Medical Information Processing, Biometry and Epidemiology, Ludwig-Maximilians University of Munich, Munich, Germany; [2]Department of Statistics Ludwig-Maximilians University of Munich, Munich, Germany; [3]Center of Integrated Protein Science Munich and Division of Clinical Pharmacology, Department of Medicine IV, Klinikum der Universität München, Munich, Germany; [4]German Center for Translational Cancer Research (DKTK), Partner Site Munich, Germany; [5]Einheit für Klinische Pharmakologie (EKLiP), Helmholtz Zentrum München, German Research Center for Environmental Health (HMGU), Neuherberg, Germany; nellenbach@ibe.med.uni-muenchen.de

There is an increasing awareness that data analysts face many uncertain choices when analyzing empirical data. These uncertain choices, which are commonly referred to as “researcher degrees of freedom”, lead to a multiplicity of possible analysis strategies that may yield overoptimistic and non-replicable research findings if combined with result-dependent selective reporting. Improvements in statistical planning and study design, such as power analyses and the pre-registration of studies, are often advocated as solutions to improve the replicability and credibility of research findings.

More specifically, appropriate power analyses and sample size calculations are essential to avoid underpowered studies, which have a higher risk of producing false negative findings and are thus more likely to yield misleading results. At the same time, a smaller sample size is often desirable, due to practical and financial aspects for example, and, in the case of preclinical animal studies or clinical studies on humans, even essential from an ethical point of view. The sample size should therefore not be larger than necessary to achieve the desired power.

However, power analyses and sample size calculations require many assumptions concerning parameters such as the expected effect size and the variability of the outcome in the target population, as well as the validity of distributional assumptions. Of course, if researchers had perfect knowledge on all parameters and assumptions, they would not need to conduct the planned study in the first place. Power analyses are thus affected by several researcher degrees of freedom, which, depending on the combination of choices, can lead to very different required sample sizes.

We discuss different researcher degrees of freedom in power analyses and sample size planning and evaluate their impact on the resulting statistical power and the required sample size. The opportunistic use of some of these researcher degrees of freedom is problematic, whereas others can be used in an unproblematic way to ensure the smallest possible sample size while still providing the study with sufficient sensitivity and validity. We illustrate these ideas using several examples from preclinical and clinical research, including a confirmatory preclinical animal study on the effectiveness of CAR T cells in tumour therapy.

Tuesday, 05/Sept/2023 12:00pm - 12:20pm
ID: 451 / **S25: 4**
Presentation Submissions - Featured Session
Featured Sessions: Young Statisticians Sessions and Panel Discussion
Keywords: antibody kinetics, within-host deterministic compartmental models, study design

Modelling antibody kinetics – A systematic review and study design considerations

Stefan Embacher, Andrea Berghold, Martin Hönigl, Sereina Herzog
Medical University of Graz, Austria; stefan.embacher@medunigraz.at

Introduction
Immunity against infectious diseases is strongly driven by antibodies. Interventions, which can reduce the burden of infections, like vaccination, need to be evaluated through clinical studies. Being able to describe antibody kinetics, the change in antibody titer over time, is crucial in optimizing the design of immunization trials. This includes determining the appropriate sample size and sampling times to accurately describe the underlying kinetics.

Methods
We established a systematic review of models used to describe antibody kinetics and how they have been used in the process of study design. We implemented and extended found models to develop a framework for answering the key research question of how individuals should be monitored in immunization trials. We took a dual approach, first fixing number of individuals and samples per person, varying sampling schedules and secondly fixing the time points and varying the number of individuals to assess accuracy and variability of the estimated parameters. Where possible we provide analytical solutions. Further, we conducted simulations to evaluate the developed framework.

Results
We found 1439 abstracts, out of which 652 full texts were screened for eligibility. In total 270 publications are eligible for data extraction. Some publications contained unclear methods or insufficient information about the model. A few publications even provided wrong solutions for the system of differential equations. We saw frequent use of basic statistical models and hardly any study design considerations. Using an implemented plasma cell model, we found that frequency and timing of sampling influences the estimates and the variability of the underlying parameters.

Conclusion
The limited use of mathematical models describing antibody kinetics, especially regarding study design, highlights the need and importance of basic research. Through our work, we aim to provide a framework, which can be actively used in practice to improve infectious disease study design.

Tuesday, 05/Sept/2023 11:00am - 11:20am
ID: 135 / S23: 1
Presentation Submissions - Regular Session (Default)
Topic of Submission: Clinical trial designs (adaptive and platform designs, external controls, …), Time-to-Event Analysis
Keywords: Clinical trial, illness-death model, multiple testing, progression-free survival, overall survival

Oncology clinical trial design based on a multistate model that jointly models progression-free and overall survival

Alexandra Erdmann[2], Jan Beyersmann[2], Kaspar Rufibach[1]
[1]Methods, Collaboration, and Outreach Group, Product Development Data Sciences, F. Hoffmann-La Roche, Switzerland;
[2]Institute of Statistics, Ulm University, Ulm, Germany; kaspar.rufibach@roche.com

When planning an oncology clinical trial, the usual approach is to assume an exponential distribution for the time-to-event endpoints. Often, besides the gold-standard endpoint overall survival, progression-free survival is considered as a second confirmatory endpoint. We use a survival multistate model to jointly model these two endpoints and find that neither exponential distribution nor proportional hazards will typically hold for both endpoints simultaneously. The multistate model approach allows us to consider the joint distribution of the two endpoints and to derive quantities of interest as the correlation between overall survival and progression-free survival. In this paper, we use the multistate model framework to simulate clinical trials with endpoints OS and PFS and show how design planning questions can be answered using this approach. In addition to the major advantage that we can model non-proportional hazards quite naturally with this approach, the correlation between the two endpoints can be exploited to determine sample size and type-I-error. We consider an oncology trial on non-small-cell lung cancer as a motivating example from which we derive relevant trial design questions. We then illustrate how clinical trial design can be based on simulations from a multistate model. Key applications are co-primary endpoints and group-sequential designs in pivotal clinical trials. Simulations for these applications show that the standard simplifying approach often leads to underpowered or overpowered clinical trials. Our approach is quite general and can be extended to more complex trial designs, further endpoints, and other therapeutic areas.

Wednesday, 06/Sept/2023 9:30am - 9:50am
ID: 449 / **S46: 4**
Presentation Submissions - Featured Session
Featured Sessions: Best practices for Data Monitoring Committees and how to get there
Keywords: Data Monitoring Committee, Data Analysis Center

Behind the scenes: the Data Analysis Center for Data Monitoring Committees

Benjamin Esterni
Cytel, France; benjamin.esterni@cytel.com

To make informed recommendations to the sponsor, Data Monitoring Committees (DMCs) periodically receive sensitive and potentially unblinded data, during a period of study conduct when the sponsor study team is still blinded. The involvement of a Data Analysis Center (DAC), independent from the Sponsor study team, is the optimal way to facilitate the preparation and presentation of the DMC reports. This team can be unblinded and can facilitate ad hoc DMC requests without compromising the integrity of the clinical trial. In this session, we will present the role of the DAC, from the receipt of trial data to the DMC's recommendations during the closed session. Different programming operational models will be described, associated with their respective advantages and challenges for the sponsor, the DMC, and the DAC, and with a particular focus on the capability for the DAC to respond to ad hoc DMC requests without involving the study team.

Monday, 04/Sept/2023 5:30pm - 5:50pm
ID: 398 / S18: 5
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), Machine Learning and Data Science, Real world data and evidence, Time-to-Event Analysis, Statistical hypothesis testing (covariate adjustment, nonparametric methods, multiple comparisons, …)
Keywords: Dataset Similarity, Propensity Scores, Oversampling Techniques, Small Data

Quantification Of Dataset Similarity For Small Sample Sizes

Maryam Farhadizadeh[1], Max Behrens[1], Angelika Rohde[2], Daniela Zöller[1]
[1]Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center - University of Freiburg, Germany, Germany; [2]University of Freiburg, Department of Mathematical Stochastics, Freiburg im Breisgau, Germany;
maryam.farhadizadeh@uniklinik-freiburg.de

Quantification Of Dataset Similarity For Small Sample Sizes:

In clinical studies, there are several situations where the research question requires identifying similarities between two datasets. For instance, when the aim is to improve the prediction model for a specific data site with limited observations, one approach is to include weighted data from external sites based on their similarity while maintaining the original distribution of the target data site. To calculate the weights and thus to quantify the similarity, one can make use of the inverse probability of belonging to the target site using logistic regression. However, the similarity may be underestimated by this approach if the target site has significantly fewer observations than the external even if the target and the external dataset are samples of the same population. To address this problem, we propose oversampling techniques.

Specifically, we have developed an iterative process where we continue to oversample the target site observations until we observe that the distributions of the target data, before and after including weighted data, remain the same. For comparing the distributions, we use Kullback-Leibler divergence and parametric methods. By carefully monitoring the distribution of the target data, we can avoid introducing bias while still increasing the sample size based on similarity. We evaluate the effectiveness of our proposed method using a simulation study under two scenarios: one with similar observations in two data sets and one with dissimilar observations with varying degrees of similarity. We compare the results obtained with and without oversampling and assess the impact of oversampling techniques on the performance in terms of prediction performance. Results indicate that oversampling the small target data can improve the quantification of similarity for obtaining weights in a prediction model, resulting in higher weights when the distributions of datasets are similar but not overestimating the weights when the datasets are dissimilar.

We demonstrate our approach using the International Stroke Trial (IST) data, including patients with acute stroke in different countries. The aim is to include weighted data from other countries to a country with limited data, while the distribution of target data plus weighted external data remains similar.

(This presentation is a joint presentation with the GMDS conference, and will also be presented at that conference with a different focus.)

Monday, 04/Sept/2023 3:20pm - 3:40pm
ID: 440 / S14: 5
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), Real world data and evidence
Keywords: Malaria, Field Trial, Bayesian hierarchical model, model selection

A randomized, double-blind placebo-control study assessing the protective efficacy of an odour-based 'push-pull' malaria vector control strategy in reducing human-vector contact

Ulrike Fillinger[1], Adrian Eugen Denz[2,3], Margaret Mendi Njoroge[1], Mohamed Mgeni Tambwe[4], Willem Takken[5], Joop J.A. van Loon[5], Sarah Jane Moore[4], Adam Saddler[4], Nakul Chitnis[2], Alexandra Hiscox[6]

[1]International Centre of Insect Physiology and Ecology, Kenya; [2]Swiss Tropical and Public Health Institute, Switzerland; [3]University of Nottingham, United Kingdom; [4]Ifakara Health Institute, Tanzania; [5]Wageningen University & Research, The Netherlands; [6]Arctech Innovation, United Kingdom; adi@adriandenz.com

Malaria is an infectious disease transmitted by Anopheles mosquitoes (the 'vector') and still kills more than half a million people globally each year. Reducing the human-vector contact ('vector control') by insecticide-treated nets and indoor residual spraying is the most effective public health measure to control malaria. While the death toll was almost halved since 2000, the global progress in malaria control has drastically slowed down since about 2015, presumably to a large extent due to gaps in vector control, such as outdoor transmission. Therefore, new vector control tools addressing outdoor transmission and settings difficult to reach with nets or spraying are urgently needed.

We implemented a randomized double-blind placebo-controlled field study in Ahero, western Kenya, to evaluate a transfluthrin-based spatial repellent ('push' intervention), an odour-baited trap ('pull' intervention), and the combined 'push-pull' package. The primary outcomes were outdoor and indoor human-vector contact, measured by human landing catches and and light-traps catches, respectively. We analysed the mosquito count data with Bayesian hierarchical, regression type models, with inference by Hamiltonian Monte Carlo (stan). After extensive model selection by leave-one-out cross-validation, we averaged two models jointly accounting for all possible dependencies by experimental design. As Bayesian data analysis isn't yet common in this field, the publication puts emphasis on carefully explaining the modelling framework.

Mosquito count data is typically very variable and highly dependent on the small-scale local environment and weather conditions. Still, similar studies usually only report interval estimates of the intervention effect with respect to an average situation (average house and week), and thus do not cover the variability of the intervention effect across different houses and weeks. We advocate for a more robust intervention assessment and therefore present a complementary arbitrary house-week analysis with interval estimates (highest density credible intervals) reflecting the intervention effect to be expected for an unknown house and week in the field.

We could demonstrate a strong protective efficacy of the spatial repellent against indoor biting but failed to achieve any protection from outdoor biting malaria vectors, by either of the three intervention strategies. The reason for adding an attractive trap to the spatial repellent was to avoid diverting bites to others, and in fact, our results indicate that repelling vectors from the indoor space resulted in an increased biting in the outdoor space. Hence there remains an urgent need to further develop and evaluate odour-baited attract-and-kill approaches that can be effectively combined with spatial repellents for a push-pull intervention for malaria control.

Wednesday, 06/Sept/2023 10:40am - 11:00am
ID: 453 / **S51: 1**
Presentation Submissions - Featured Session
Featured Sessions: IBS-DR/ROeS Award session

Online multiple testing with FWER control

Lasse Fischer
University of Bremen, Germany; fischer1@uni-bremen.de

While online FDR control is studied extensively, there is less work on FWER control in an online multiple testing setting. In 2021, Tian & Ramdas introduced the Adaptive-Discard-Spending (ADDIS-Spending) as online procedure with FWER control. In my master's thesis we built on this and developed the ADDIS-Graph, which combines the ADDIS concept with the graphical approach by Bretz et al. (2009). In addition to easier interpretability, our ADDIS-Graph leads to a higher power in the case of locally dependent p-values and asynchronous testing setups than the ADDIS-Spending. We further exhausted the significance level under independence of the p-values to obtain uniformly superior ADDIS procedures. Moreover, we formulated a new closure principle for online multiple testing and derived a condition under which a closed procedure is indeed an online procedure. In this talk, I summarise our theoretical findings that are supported by simulations.

Wednesday, 06/Sept/2023 9:30am - 9:50am
ID: 137 / **S44: 3**
Presentation Submissions - Invited Session
Invited Sessions: Online hypothesis testing and subgroup analyses in complex innovative designs
Keywords: graphical testing procedures, false discovery rate, familywise error rate, online multiple testing.

Graphical procedures for online error control

Lasse Fischer[1], **Marta Bofill Roig**[2], **Werner Brannath**[1]
[1]University of Bremen; [2]Medical University of Vienna; fischer1@uni-bremen.de

Bretz et al. (2009) proposed the representation of multiple testing procedures by directed graphs, where the null hypotheses are represented by nodes accompanied by their individual significance levels and connected by weighted vertices, illustrating level distribution in case of a rejection. Such graphical procedures are becoming increasingly popular, since they make the calculation of individual significance levels easy to follow and thus facilitate communication with users. In addition, graphical procedures are often very general and therefore include several other procedures. In many contemporary applications, hypotheses are tested in an online manner. This means, the hypotheses are tested one at a time without access to the future hypotheses and decisions. At each step, a type I error rate, such as familywise error rate (FWER) or false discovery rate (FDR), shall remain under control. In this presentation, we focus on the construction of graphical procedures providing error control in the online setting.

The extension of the classical graphical procedure by Bretz et al. (2009) to the online framework has been recently proposed. In this context, the graph illustrates the significance level distribution over time. However, previous work has shown that this Online-Graph leads to low power when the number of hypotheses is large, which is why other approaches, such as Adaptive-Discard (ADDIS) procedures (Tian and Ramdas, 2019, 2021), are preferred in the online setting.

In this talk, we present a new online procedure that combines the concepts of adaptivity and discarding to the Online-Graph. The resulting ADDIS-Graph controls the FWER when the p-values are independent. We show that it can be also adapted to a local dependence structure and an asynchronous testing setup, resulting in power superiority over current state-of-art methods. Furthermore, we extend the approach to construct an FDR-ADDIS-Graph with similar advantages. We illustrate the gain in power when using the ADDIS-Graph compared to previous procedures through a simulation study. The combination of easy interpretability and high online power makes these graphical ADDIS approaches well suited for a multitude of online applications, including complex trial designs, such as platform trials, but also large-scale test designs, such as those faced in genomics research.

Wednesday, 06/Sept/2023 10:40am - 11:00am
ID: 372 / **S52: 1**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Machine Learning and Data Science
Keywords: Random forests, variable selection, informative variable, important variable

Challenge in distinguishing important from informative variables in random forest prediction models

Césaire J. K. Fouodo, Sike Szymczak
Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck, Germany; cesaire.kuetefouodo@uni-luebeck.de

Random forest (RF) is a well-performing prediction method for high-dimensional data and enables the selection of predictor variables using variable importance measures. The Actual Impurity Reduction (AIR) measure is a computationally efficient and unbiased RF importance measure. Although many RF variable selection procedures are based on importance measures, how to interpret the resulting measurements in relation to the structure of the data and the prediction model is rarely questioned. In most cases, the importance of each variable is interpreted as its ability to improve the model prediction performance and, therefore, analyzed independently from the proportion of the available predictors associated with the response variable. For example, having a large proportion of associated predictors with the response variable in a dataset does not necessarily mean that all of them are important for building a predictive model.

We propose to distinguish important from informative predictor variables. A predictor variable is called informative if it is associated with the response variable. An important variable is an informative variable that substantially improves the model's prediction performance. Therefore, an informative variable can be unimportant if it does not significantly improve the predictive model. Such an unimportant informative predictor variable may be interpreted as a noise variable, although it is associated with the response variable.

We used simulation studies to demonstrate the effects of the proportion of informative variables on the estimated AIR importance of RF. We simulated datasets with non-informative noise variables and different proportions of non-correlated informative predictor variables.

Our results show that estimated AIR decreases when the proportion of informative variables in the dataset increases. We explain why this decrease in the estimated importance can strongly affect variable selection testing procedures. Finally, we expect this study to improve the interpretation of RF variable importance measures.

Wednesday, 06/Sept/2023 8:30am - 8:50am
ID: 447 / **S46: 1**
Presentation Submissions - Featured Session
Featured Sessions: Best practices for Data Monitoring Committees and how to get there
Keywords: Clinical trials, ethics, data monitoring, adverse events

Introduction to DMCs: basic principles and some statistical issues

Tim Friede
Universitätsmedizin Göttingen, Germany; tim.friede@med.uni-goettingen.de

A Data Monitoring Committee (DMC) is one out of several oversight groups monitoring the progress of a clinical trial; others include steering, adjudication or ethics committees. The talk sets out by describing situations requiring a DMC and then introduces some elementary aspects such as a DMC's composition and organization (Herson, 2009; Ellenberg et al, 2019). Central to the organization of a DMC is the Charter as it defines the meeting format including open and closed sessions, the meeting intervals, the statistical analyses to be conducted and their reports. The information flow including the DMC's interactions with the statistical analysis centre and the sponsor representative will be considered, in particular in contexts such as adaptive designs and programme level DMCs. Furthermore, we will discuss some statistical issues relevant to the monitoring of data such as the use of stopping boundaries (Mütze and Friede, 2020), the estimation of adverse event risks (Stegherr et al, 2021) and the integration of data across clinical trials (Friede et al, 2017).

References

1. Ellenberg SS, Fleming TR, Demets DL (2019) Data monitoring committees in clinical trials: a practical perspective. Wiley, 2nd edition.
2. Friede T, Röver C, Wandel S, Neuenschwander B (2017) Meta-analysis of few small studies in orphan diseases. Research Synthesis Methods 8: 79–91.
3. Herson J (2009) Data and safety monitoring committees in clinical trials. Chapman & Hall / CRC, Boca Raton, FL.
4. Mütze T, Friede T (2020) Data monitoring committees for clinical trials evaluating treatments of COVID-19. Contemporary Clinical Trials 98: 106154.
5. Stegherr R, Schmoor C, Beyersmann J, Rufibach K, Jehl V, Brückner A, Eisele L, Künzel T, Kupas K, Langer F, Leverkus F, Loos A, Norenberg C, Voss F, Friede T (2021) Survival analysis for AdVerse events with VarYing follow-up times (SAVVY) - estimation of adverse event risks. Trials 22: 420.

Thursday, 07/Sept/2023 8:50am - 9:10am
ID: 136 / S60: 2
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), High dimensional data, genetic and x-omics data
Keywords: High dimensional data, L0 penalties, Neutral comparison, Variable selection

A neutral comparison of algorithms to minimize L0 penalties for high-dimensional variable selection

Florian Frommlet
Medical University Vienna, Austria; Florian.Frommlet@meduniwien.ac.at

Variable selection methods based on L0 penalties have excellent theoretical properties to select sparse models in a high-dimensional setting. There exist modifications of BIC which either control the family wise error rate (mBIC) or the false discovery rate (mBIC2) in terms of which regressors are selected to enter a model. However, the minimization of L0 penalties comprises a mixed integer problem which is known to be NP hard and therefore becomes computationally challenging with increasing numbers of regressor variables. The last few years have seen some real progress in developing new algorithms to minimize L0 penalties. Simulation studies covering a wide range of scenarios which are inspired by genetic association studies as well as a real data example concerned with eQTL mapping are used to compare the performance of some of these algorithms. The study results in a clear recommendation which algorithms to use in practice.

Wednesday, 06/Sept/2023 9:30am - 9:50am
ID: 360 / S43: 4
Presentation Submissions - Regular Session (Default)
Topic of Submission: Estimands and causal inference
Keywords: causal inference, overlap weighting, double machine learning, target trial

Estimating and interpreting causal effects under violation of positivity

Maria Geers[1], **Vanessa Didelez**[1,2]
[1]Leibniz Institute for Prevention Research and Epidemiology – BIPS, Germany; [2]University of Bremen, Germany;
geers@leibniz-bips.de

A violation of the fundamental positivity assumption in causal inference leads to lack of overlap in the data and poses a challenge to the interpretation and estimation of causal effects. Not only for statistical reasons, but also for interpretation, the treatment strategies should, in principle, be compared only for those units for whom both strategies are possible and sensible. In a target trial emulation (TTE) this would be ensured by carefully chosen eligibility criteria. However, this choice is not always obvious. An automated way of dealing with positivity violations would be desirable, though it is not clear whether and to what extent automatization is possible. Some methods of estimation provide a more or less automated approach to lack of overlap, e.g. double machine learning or overlap weighting and other propensity score weighting methods using balancing weights; moreover, the method of entropy balancing optimizes the balance (more accurate: some predefined balance constraints) directly by choosing a suitable set of weights and propensity score (PS) matching addresses the problem by pruning unmatched observations. Often it is not clearly emphasized that these different approaches implicitly modify the estimand and/or the population, which may then differ from the originally intended ones. Moreover, especially when using software packages based on double machine learning algorithms, it is sometimes not easy to see how problems due to positivity issues are handled. In addition to the estimation of causal effects under violation of positivity, another difficulty concerns diagnostics for lack of overlap which often involve a subjective assessment of PS plots (which in turn depend on the chosen model for the PS). These diagnostics should also give hints to the cause of the lack of overlap (e.g. relevant covariates/subgroups).

In this work, we provide a general comparison of approaches and methods of estimation for causal effects regarding the estimands, the population and the handling of positivity violations; we further review and compare techniques to diagnose lack of overlap. To clarify the definition of the various estimands we use Single-World Intervention Graphs (SWIGs). These graphs are able to display the counterfactual (in)dependencies under a specific intervention. We illustrate our conclusions with semi-simulated data using the Rotterdam breast cancer dataset to estimate the causal effect of a hormonal therapy after a breast cancer diagnosis, where considerable overlap issues exist.

Our results provide guidance on the strengths and limitations of the different methods in practical applications. They also show that some of the methods are not comparable in a strict sense as they target different estimands. Finally, using a TTE framework, we discuss what aspects of causal effect estimation under lack of positivity may or may not lend themselves to an automatization and where expert input is required.

Monday, 04/Sept/2023 12:20pm - 12:40pm
ID: 170 / **S6: 5**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical hypothesis testing (covariate adjustment, nonparametric methods, multiple comparisons, …)
Keywords: Epidermolysis Bullosa Simplex; Generalized pairwise comparison (GPC); Neutral comparison; Nonparametric marginal model (nparLD); Repeated measures

Comparing statistical methods for analyzing longitudinally measured ordinal outcomes in rare disease settings.

Martin Geroldinger[1,2], Johan Verbeeck[3], Konstantin E. Thiel[1,2], Geert Molenberghs[3,4], Arne C. Bathke[5], Martin Laimer[6], Georg Zimmermann[1,2]

[1]Team Biostatistics and Big Medical Data, IDA Lab Salzburg, Paracelsus Medical University Salzburg, Austria; [2]Department of Research and Innovation, Paracelsus Medical University, Salzburg, Austria; [3]Data Science Institute (DSI), Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BioStat), Hasselt University, Belgium; [4]Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BioStat), KULeuven, Belgium; [5]Intelligent Data Analytics (IDA) Lab Salzburg, Department of Artificial Intelligence and Human Interfaces, Faculty of Digital and Analytical Sciences, Paris Lodron University of Salzburg, Austria; [6]Department of Dermatology and Allergology, Paracelsus Medical University, Salzburg, Austria;
martin.geroldinger@pmu.ac.at

Ordinal data in a repeated measures design of a cross-over study for rare diseases usually do not allow for the use of standard parametric methods. Hence, nonparametric methods should be considered instead. Determination of an appropriate nonparametric approach is likewise challenging, as only limited simulation studies for complex trial designs with very small sample sizes exist. Referring to a cross-over trial for the genodermatosis epidermolysis bullosa, a rank-based approach using the R package nparLD and different generalized pairwise comparisons (GPC) methods were assessed neutrally in a comparative simulation study. The results revealed no single best method for this particular design, since a trade-off became apparent between achieving high power, accounting for period effects, and controlling for missing data. Specifically, nparLD as well as unmatched GPC approaches did not address cross-over aspects, and the univariate GPC variants partly ignored longitudinal information. The matched GPC approaches, on the other hand, took the cross-over effect into account in the sense of incorporating the within-subject association. Overall, the prioritized unmatched GPC method achieved the highest power in the simulation scenarios, although this may be due to the specified prioritization. The rank-based approach yielded good power even at a sample size of N=6, while the matched GPC method could not control the type I error. Together with the results from extensive simulation studies using binary and count outcome data, our findings will add to the development of recommendations and educational materials which will be disseminated in the statistical as well as in the clinical scientific community. Thereby, the accurateness of methodological approaches of clinical research in rare diseases should be increased.

(This research has been conducted within the framework of the EBStatMax project, which is funded by the European Joint Programme on Rare Diseases, EU Horizon 2020 grant no. 825575. The Authors gratefully acknowledges the support of the WISS 2025 project 'IDA-Lab Salzburg' (20204-WISS/225/197-2019 and 20102-F1901166-KZP).)

Tuesday, 05/Sept/2023 3:20pm - 3:40pm
ID: 329 / **S29: 5**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Generalized pairwise comparisons
Keywords: Generalized pairwise comparisons, Net Benefit, N-of-1 trials, meta-analyses, personalized medicine

Individualized Net Benefit estimation and meta analysis using generalized pairwise comparisons in N-of-1 trials

Joris Giai[1,2], Julien Péron[2,3,4], Matthieu Roustit[5], Jean-Luc Cracowski[5], Pascal Roy[2,3], Brice Ozenne[6,7], Marc Buyse[8,9], Delphine Maucort-Boulch[2,3]

[1]Univ. Grenoble Alpes, Inserm CIC1406, CHU Grenoble Alpes, TIMC UMR 5525, 38000 Grenoble, France.; [2]Université de Lyon, Université Lyon 1, CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Évolutive, Équipe Biostatistique-Santé, Villeurbanne, France.; [3]Hospices Civils de Lyon, Pôle Santé Publique, Service de Biostatistique - Bioinformatique, Lyon, France.; [4]Hospices Civils de Lyon, Oncology department, Pierre-Bénite, France.; [5]Univ. Grenoble Alpes, Inserm CIC1406, CHU Grenoble Alpes, HP2 Inserm U1300, 38000 Grenoble, France.; [6]Neurobiology Research Unit, Rigshospitalet, Copenhagen, Denmark; [7]University of Copenhagen, Department of Public Health, Section of Biostatistics, Copenhagen, Denmark; [8]International Drug Development Institute (IDDI), San Francisco, CA, USA; [9]Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-Biostat), Hasselt University, Hasselt, Belgium; JGiai1@chu-grenoble.fr

Background: The Net Benefit (Δ) is a measure of the benefit-risk balance in clinical trials, based on Generalized Pairwise Comparisons (GPC) using several prioritized outcomes and thresholds of clinical relevance. We extended Δ to N-of-1 trials, with a focus on patient-level and population-level Δ.

Methods: We developed a Δ estimator at the individual level as an extension of the stratum-specific Δ, and at the population-level as an extension of the stratified Δ. We performed a simulation study mimicking PROFIL (NCT02050360), a series of 38 N-of-1 trials testing low and high-dose sildenafil versus placebo in Raynaud's phenomenon on three outcomes, to assess the power for such an analysis with realistic data. We then reanalyzed PROFIL using individual-level GPC with expert-defined outcome hierarchy and validated minimal clinically important differences (MCID) acting as thresholds of clinical relevance. This reanalysis was interpreted in the context of the main analysis of PROFIL which was performed in a Bayesian framework and reported : i) individual probabilities of efficacy, and ii) individual adjusted risk variations. We showed a straightforward way to aggregate individual-level Δ in order to estimate a population-level Δ, while highlighting similarities and differences between our aggregation method and a more usual random-effects meta-analysis.

Results: Simulations under the null showed good size of the test for both individual and population levels. The test lacked power when being simulated from the true PROFIL data, even when increasing the number of repetitions up to 140 days per patient. PROFIL individual-level estimated Δ were well correlated with the probabilities of efficacy from the Bayesian analysis while showing similarly wide confidence intervals. Likewise, individual-level Δ led to similar conclusions than individual adjusted risk variations. Population-level estimated Δ was not significantly different from zero, consistently with the previous Bayesian analysis.

Conclusion: GPC can be used to estimate individual Δ which can then be aggregated in a meta-analytic way in N-of-1 trials. We argue that GPC ability to easily incorporate patient preferences (thresholds of clinical relevance on the same scale as the outcome itself and outcome prioritization) allow for more personalized treatment evaluation, while needing much less computing time than Bayesian modeling. Finally, we discuss the current limits of GPC usage in N-of-1 trials and some ways to alleviate them, as well as undergoing developments.

Monday, 04/Sept/2023 2:00pm - 2:20pm
ID: 274 / **S11: 1**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Safety and benefit/risk assessment in master protocols
Keywords: safety analysis, platform trials, estimands, competing risks

Analysis of safety data with special attention to platform trials and the estimand framework

Ekkehard Glimm
Novartis Pharma AG, Switzerland; ekkehard.glimm@novartis.com

The analysis of safety data poses many statistical challenges, especially when done on clinical trials that had been designed with a focus on efficacy. Sample sizes are often inadequate for conclusive statements about rare, but serious events, imposing the need for cross-study comparisons. Furthermore, serious adverse of events of special interest will often require closer scrutiny than routinely collected, less serious adverse events (AEs). Hence, both data collection and analyses methods might differ for these two types of AEs.

In this talk, we will give an overview of methods which can be applied to the analysis of AEs. Special attention will be given to the fact that the assessment of safety risks often has to rely on non-randomized comparisons, which complicates causal interpretations. In platform trials, for example, safety data on different treatments will be collected non-concurrently, requiring the handling of potential time trends in the emergence of safety signals.

Duration of exposure to a risk (e.g. a treatment) is also an important component of the risk assessment. Exposure to risk might differ between treatment arms and between studies. In addition, its observation can be masked by intercurrent events such as withdrawal of a patient from the study or by the occurrence of AEs, including other AEs which prevent observation of the AE of interest. A decision then must be made if the intercurrent event induces censoring or if it should be viewed as a competing risk.

Depending on the importance of time and the presence of competing risks, simple estimates of event probabilities in a given time frame, estimates of incidence rates, or cumulative incidence functions might be the most appropriate summary of risk. Regarding inference, available tools range from meta-analyses of crude event probabilities via Cox-regression models for time to AE and models for restricted mean survival times, for the cause-specific or the subdistribution hazard to multistate models.

To aid the decision about the most appropriate analysis tool, it is important to obtain clarity about the safety estimand which is targeted. As an example, the risk of (non-lethal) strokes may appear increased relative to untreated patients by a treatment which reduces the probability of dying from a stroke, but does not influence the frequency of strokes. Hypothetical and composite estimands can be considered, but have different interpretations and require different analysis methods.

The talk will illustrate possible approaches to dealing with these topics with examples from various therapeutic areas.

Tuesday, 05/Sept/2023 11:40am - 12:00pm
ID: 334 / S27: 3
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Statistical issues in health care provider comparisons
Keywords: PROM repeated measures, intercurrent events, quality of care

Analysing PROM based quality of care indicators in care centers

Els GOETGHEBEUR
Ghent University, Belgium; els.goetghebeur@ugent.be

In clinical trials as in the evaluation of patient care in care centers, the importance of quality of life besides more standard clinical and process outcomes has increasingly been recognised. Among the challenges of the assessment of repeated PROM measures in older people we recognise that 1) death may enter as an endpoint while no values post death can be considered that have concrete meaning; 2) death may be preceded by a period of substantial decline in QOL possibly accompanied by missing data; 3) not only before death, but also after moving to a care center one may see a temporary shift in QOL and 4) QOL measures are possibly subject to a response shift over time, irrespective of changes in the physical care environment.

We discuss several estimands in this light, focusing on QOL-while-alive combined with survival and how it can be meaningfully compared within and between care centers. Following the estimand framework guidance we seek to define the population (involving a core set of baseline covariates), the treatment (center) and outcome (following 3-monthly PROM measures, say) along with a relevant summary statistic to be meaningfully compared between centers. The latter must handle `intercurrent events' such as death and possibly hospitalization or other disruptive phases in nursing home care. In function of the (multi-dimensional) estimand several estimators are considered. Mixed models, for instance, may be useful to model observed longitudinal data but cannot be implemented naively to avoid latent imputation of QOL measures after death.

Reference:

Brenda F. Kurland, Laura L. Johnson, Brian L. Egleston and Paula H. Diehr. `Longitudinal Data with Follow-up Truncated by Death: Match the Analysis Method to Research Aims'. Statistical Science 2009, Vol. 24, No. 2, 211–222

Wednesday, 06/Sept/2023 9:50am - 10:10am
ID: 456 / **S46: 5**
Presentation Submissions - Featured Session
Featured Sessions: Best practices for Data Monitoring Committees and how to get there
Keywords: Data monitoring committee, sponsor perspective

Sponsor perspective on best practices for Data Monitoring Committees

Gregory T Golm[1], Tobias Muetze[2], Harald Siedentop[3]
[1]MSD, United States of America; [2]Novartis; [3]Bayer; gregory_golm@merck.com

The sponsor of a study establishes the DMC and sets up the structure by which the DMC and the independent statistical analysis center operate, described in the DMC charter. This presentation will highlight best practices for the sponsor to enable the DMC and the independent statistical analysis center to function efficiently, with an emphasis on ensuring study integrity. The presentation will also suggest areas for improvement in the DMC process from the sponsor perspective. Considerations for forming the DMC will be discussed, and the importance of prespecifying operating procedures and communication channels will be emphasized.

Tuesday, 05/Sept/2023 4:10pm - 4:30pm
ID: 389 / **S41: 1**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Clinical trial designs (adaptive and platform designs, external controls, …), Real world data and evidence, Time-to-Event Analysis
Keywords: early drug development, estimation, external control data, time-to-event data

Estimation of treatment effects in early phase randomized clinical trials involving external control data

Heiko Götte[1], Marietta Kirchner[2], Johannes Krisam[2], Arthur Allignol[3], Armin Schüler[1], Meinhard Kieser[2]
[1]Merck Healthcare KGaA, Germany; [2]Institute of Medical Biometry, University of Heidelberg, Germany; [3]Daiichi Sankyo Europe GmbH; heiko.goette@merckgroup.com

Randomized controlled trials (RCT) are the gold standard design even for early phase development as they provide unbiased treatment effect estimates. However, the low sample sizes in those settings lead to high variability of the treatment effect estimate. This variability could be reduced by adding external control data (augmented RCT) which might introduce bias in return. For the common setting of suitable subject-level control group data only available from one external (clinical trial or real-world) data source, we evaluate different analysis options for estimating the marginal treatment effect in the target, i.e. RCT, population via hazard ratios. The analyses options have in common that the contribution of the external control data is usually driven by the level of similarity with the current RCT data. Such level of similarity can be assessed via outcome and/or baseline covariate data comparisons. We provide an overview over existing methods where we focus on Bayesian hierarchical models (BHM) for the outcome and propensity-score (PS) based approaches for baseline covariates.

We propose a novel option which includes an outcome and a baseline data component: a model averaging-based combination of BHM estimates and PS-based estimates. The relative contribution of the estimates is determined by a model averaging weight which reflects the Likelihood contributions of RCT subjects after fitting BHM and PS-models to the complete data set. The reasoning behind this approach is twofold: the different ways in handling external data makes the comparison of the Likelihood contributions of external subjects between BHM and PS-models difficult and better model fit just due to the external control subjects would have a higher potential of introducing bias.

In our simulation study under varying assumptions regarding observable and unobservable confounder distributions using a time-to-event model, we compare a selection of existing methods as well as the proposed approach. Our various simulation scenarios also reflect the differences between external clinical trial and real-world data.

One of our main findings is that only very few analysis options for augmented RCTs perform better than using the RCT estimate which ignores the external control data. “Better” means substantially lower variability (by means of root mean square error (RMSE)) with low bias across various scenarios. The non-favorable options either introduce too much bias in certain scenarios (simple PS weighting), have limited RMSE reduction with moderate bias (unadjusted BHM) or increased RMSE (PS matching before fitting hierarchical model) compared to the RCT estimate. Only two analysis options which conflate outcome and baseline covariate data perform better than the RCT estimate: a marginalized estimate from a covariate adjusted hierarchical model (AHM) and our proposal, which might use AHM as its outcome component.

Monday, 04/Sept/2023 4:30pm - 4:50pm
ID: 303 / **S21: 2**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …)
Keywords: estimator, optimality criteria, weighing design

How to estimate parameters in weighing design?

Małgorzata Graczyk
Poznań University of Life Sciences, Poland; malgorzata.graczyk@up.poznan.pl

Here, the issues related to the estimation of parameters in the model of spring and chemical balance weighing designs are presented. These problems are discussed from the point of view of different optimality criteria and different assumptions regard to the errors. We consider the models with uncorrelated errors with different variances and equally correlated errors. The forms of estimators and the necesseary and sufficient conditions determining them are given.

Monday, 04/Sept/2023 3:20pm - 3:40pm
ID: 415 / S10: 5
Presentation Submissions - Regular Session (Default)
Topic of Submission: Biomarkers and diagnostics, Statistical modelling (regression modelling, prediction models, …), Personalized health care, High dimensional data, genetic and x-omics data
Keywords: prediction modelling, network inference, personalized medicine, flexible parametrization, plasmode simulation

Individual-specific network inference for prediction modelling: a plasmode simulation study

Mariella Gregorich, **Georg Heinze**
Medical University of Vienna, Austria; mariella.gregorich@meduniwien.ac.at

Statistical techniques are needed to analyse data structures with complex dependencies such that clinically useful information for personalized medicine can be extracted. Individual-specific networks involve the estimation of a separate connectivity matrix (adjacency matrix) across a common set of nodes for each individual subject and can provide graph-theoretical features for prognostic outcome modelling. In particular, neuroimaging studies have demonstrated the potential of network connectivity patterns estimated from functional magnetic resonance imaging (fMRI) to discriminate between diagnostic groups. Despite the growing use of graph theory in these studies, many techniques for network inference require an edge weight thresholding procedure that comes with uncertainties when selecting the appropriate threshold. This can lead to a broad range of network representations for a single individual, resulting in a high degree of variation in the extracted graph-theoretical features.

We propose a flexible parametrization approach that accounts for the full range of possible thresholds and their corresponding graph-theoretical features. This is achieved by fitting a weight function (e.g. using penalized splines) that assigns a weight to the feature at each threshold based on its relative leverage on the outcome and is determined using statistical goodness-of-fit criteria, while also accommodating structural constraints of the function. By doing so, it enables us to incorporate uncertainties in individual-specific network inference in the model and provides greater flexibility in network sparsification. We conducted a plasmode simulation study using preprocessed fMRI data (N=686) from the Autism Brain Imaging Data Exchange (ABIDE) initiative (1) to provide evidence for a proof-of-concept of our proposed methodology and (2) to obtain semi-simulated data that inherits the complex data structure derived from fMRI data and their variability across individuals. Under various data-generating conditions, we compare the prognostic performance of our methodology against current state-of-the-art methods, which comprise parameter space sampling to select the optimal threshold and averaging across graph-theoretical features corresponding to a subset of thresholds to avoid one potentially ill-informed threshold. The estimands in the simulation study are the functional relationship of the network feature-outcome association and the prognostic performance of models. Performance is assessed by the cross-validated average root mean squared error and the R^2 of the predictions.

The new flexible approach presents noticeable advantages over the current state-of-the-art methods across various data-generating conditions. Specifically, when dealing with larger sample sizes and distinct noise scenarios that impact network structure in real-life applications, the flexible approach demonstrated superior performance. Nonetheless, comprehensive simulation studies are required to evaluate the impact and uncertainty of typical network inference approaches We demonstrate that flexible parametrization of graph-theoretical features can be a valuable tool for prediction, provided that researchers are mindful of the intricate aspects of parameter tuning and their behavior in diverse contexts. We highlight certain challenges that must be overcome before our approach can be routinely applied and provide suggestions when using our proposed approach in practical data settings.

Wednesday, 06/Sept/2023 10:40am - 11:00am
ID: 122 / **S55: 1**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Clinical trial designs (adaptive and platform designs, external controls, …), Preclinical drug development, safety and toxicology
Keywords: consultancy; historical data; mathematical risk assessment; acute toxicity

"Lots of time to think": Statistical Consultancy at Ciba-Geigy in the Early 1980s

Andrew Grieve
UCB Celltach, United Kingdom; Andy.Grieve@UCB.com

In 1979 I joined the Mathematische Applikationen section of Ciba-Geigy's Wissenschaftliches Rechenzentrum in Basel. I was mainly active in providing support to the toxicology functions in both Ciba-Geigy's agrochemical and pharmaceutical businesses, but had the space and time to investigate the use of Bayesian methods broadly within Ciba-Geigy's various businesses. In this talk I will look at some of the statistical applications that I was involved in from determining the safe level of Urethane in kirsch for the Bundesamt für Gesundheit in Bern, to providing evidence that batches of a pesticide was no more toxic than Ciba-Geigy's published data as claimed by one national authority. I will also address approaches to the use of historical data, a topic of continuing interest today, and the appropriate experimental units in one class of toxicology experiments which has links to cluster randomised trials.

Wednesday, 06/Sept/2023 10:40am - 11:00am
ID: 197 / **S53: 1**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Free Contributions
Keywords: EU HTA, EUnetHTA 21

EUnetHTA 21 methods guidelines for EU HTA: The good, the bad, and the ugly - an industry perspective.

Sandro Gsteiger, Maximo Carreras, Stefanie Hieke-Schulz, Kaspar Rufibach, Alex Simpson, Hannah Staunton, Anna Steenrod, Lutz Westermann

F. Hoffmann-La Roche Ltd.; sandro.gsteiger@roche.com

In less than 2 years, oncology and advanced therapy medicinal products will undergo the joint European HTA process defined in the EU HTA Regulation (HTAR) adopted by the European Parliament in Dec 2021 [1]. EUnetHTA 21, a consortium of 13 national HTA organizations, has developed proposals for methodological guidelines as part of a service contract for the European Commission [2]. The final methodological guidelines will be adopted by the EU HTA Coordination Group, composed by representatives of all 27 Member States, before January 2025. However, some uncertainty remains regarding the final methods for European HTA, though they will likely rely heavily on the proposals developed by EUnetHTA 21.

From a pharmaceutical industry perspective, we consider that the EUnetHTA 21 documents provide a mixed picture. Positive elements include the clear discussion of the types of clinical outcome assessments (in line with for example the FDA categorization of COAs), the description of “classical” evidence synthesis methods such as pairwise meta-analysis and (frequentist and Bayesian) network meta-analysis, and the inclusion of the estimands framework (at least in parts). Nevertheless, there are also several shortcomings with the proposed guidelines. These include the critical view on the role of non-randomised evidence in decision making, the lack of methods such as target trial emulation or the use of external controls, the problem of how to address the multiplicity of PICOs, and the relatively large focus on hypothesis testing.

Overall, we see a risk that the proposed guidelines by EUnetHTA 21 fail to increase the level of harmonization of methods and approaches used for the clinical assessment of health technologies. The reduction of duplication and of fragmentation of the HTA landscape - sought by the EU HTAR - may not be achieved, which may in turn harm patients who rely on timely access to innovative therapies. In our view, methodological guidelines should seek to establish a common understanding of a pan-EU HTA approach. The scientific foundations of Health Technology Assessment should be harmonized more than what is currently proposed by EUnetHTA 21. Without this directional change in the methodological guidelines, the novel EU Joint Clinical Assessment may add an additional layer of complexity instead of reducing duplication and, therefore, may fail to improve patient access in Europe.

We will shortly introduce the key pillars of the EU HTAR, summarize the main elements of the EUnetHTA 21 methodological guideline proposals, point out what - in our view - are the main issues from a pharmaceutical industry perspective, propose some of the directional changes needed to achieve the objective of faster and sustainable patient access set out by the EU HTAR, and elaborate on the role statisticians can play in this process.

References:

[1] European Commission. “Regulation (EU) 2021/2282 of the European Parliament and of the Council of 15 December 2021 on Health Technology Assessment and Amending Directive 2011/24/EU (Text with EEA Relevance),” 458 OJ L § (2021), http://data.europa.eu/eli/reg/2021/2282/oj/eng.

[2] EUnetHTA. Joint HTA Work. https://www.eunethta.eu/jointhtawork/.

Monday, 04/Sept/2023 12:00pm - 12:20pm
ID: 308 / **S6: 4**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), High dimensional data, genetic and x-omics data

Methods of Model selection for models with common parameters

Onur Gül, Kirsten Schorning
TU Dortmund, Germany; onur.guel@tu-dortmund.de

The analysis of gene-expression data leads to a high-dimensional statistical problem where thousands of concentration-response data have to analysed. For instance, the concentration-response data provided in the Valproic acid (VPA) data set the information about the concentration-response relationship of more than 20.000 genes. Fitting each of these concentration-response data separately to a non-linear model leads to a complex model with many parameters and a corresponding high-dimensional estimator with high variance.

Assuming that some genes behave similarly and that the corresponding concentration-response data can be fitted by non-linear models with common parameters, can reduce the number of unknown parameters substantially. In particular, it might be reasonable that the concentrations at which 50% of the maximum effect is achieved (EC_{50}) are at least similar for some genes and therefore these parameters can be assumed to be the same across the considered non-linear models. This assumption causes a reduction of the variance of the lower-dimensional parameter estimator, but also a bias, as the assumed shared parameters are only similar, but not the same.

In this talk, we answer the question under which circumstances the less complex model with the additional assumption of common parameters should be used instead of the complex model where all genes are considered separately. More precisely, we derive asymptotic properties of the estimators in each of the models in order to calculate the asymptotic mean squared errors. Based on the asymptotic, we derive a model selection criterion which selects the model (with common parameters) leading to the smallest mean squared error.

We show in a simulation study that the derived model selection criterion performs well in comparison to other common selection criteria. Moreover, we apply the developed model selection criterion to the VPA data set in order to estimate the EC_{50}.

Thursday, 07/Sept/2023 9:10am - 9:30am
ID: 278 / **S59: 3**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Volume-outcome relationships in health care
Keywords: health care quality measurement, volume-outcome analysis, minimum provider volume, additive regression models

Modelling volume-outcome relationships in health care

Maurilio Gutzeit, Johannes Rauh, Jona Cederbaum
IQTIG, Germany; maurilio.gutzeit@iqtig.org

Despite the ongoing strong interest in associations between quality of care and the number of cases (volume) of healthcare providers, a unified statistical framework for analyzing them is missing. Also, many studies suffer from poor statistical modelling choices such as the discretization of volume into groups.

In this talk, we present a flexible, additive mixed model on the level of the individual patients for studying volume-outcome associations in health care. We treat volume as a continuous variable and model its effect on the considered outcome through penalised splines. We adjust for different case-mixes by including patient-specific risk factors. Furthermore, we take into account clustering on the provider level through random intercepts. Using that approach, we obtain a smooth volume effect as well as volume-independent provider effects. A comparison of these two quantities gives insight into the sources of variability of quality of care. All effects are estimated in a unified framework allowing for adequate uncertainty quantification.

Depending on the estimated association from data, our approach also enables the estimation of potential threshold values for the volume based on a break point model. For instance, that is of interest when investigating administrative requirements on the minimum provider volume. Furthermore, given a potential minimum provider volume, it is also possible to evaluate the statistical effect on the number of adverse outcomes.

We illustrate our approach through an example based on German health care data.

Thursday, 07/Sept/2023 12:00pm - 12:20pm
ID: 378 / **S69: 5**
Presentation Submissions - Regular Session (Default)
Topic of Submission: High dimensional data, genetic and x-omics data
Keywords: single-cell RNA-sequencing, time-series data, dimension reduction, visualization, deep learning

More than meets the eye: Dimension reduction and temporal patterns in time-series single-cell RNA-sequencing data

Maren Hackenberg, Laia Canal Guitart, Harald Binder
Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center -- University of Freiburg, Germany;
maren.hackenberg@uniklinik-freiburg.de

Generating single-cell RNA-sequencing (scRNA-seq) data at several time points, e.g., during a developmental process, promises insights into mechanisms controlling cellular differentiation at the level of individual cells. As there is no one-to-one correspondence between cells at different timepoints, a first step in a typical analysis workflow is to reduce dimensionality to visually inspect temporal patterns. Here, one implicitly assumes that the resulting low-dimensional manifold captures the central gene expression dynamics of interest. Yet, commonly used techniques are not specifically designed to do so and their representations do not necessarily coincide with the one that best reflects the actual underlying dynamics.

We thus investigate how visual representations of different temporal patterns in time-series scRNA-seq data depend on the choice of dimension reduction, considering principal component analysis (PCA), t-distributed stochastic neighbourhood embedding (t-SNE), uniform manifold approximation and projection (UMAP) and single-cell variational inference (scVI), a popular deep learning-based approach.

To characterize the approaches in a controlled setting, we create an artificial time series from a snapshot scRNA-seq dataset by simulating an underlying low-dimensional developmental process and generating corresponding high-dimensional gene expression data. Specifically, we apply a specific dimension reduction approach (say, tSNE) on the snapshot data and transform the low-dimensional representation according to biologically meaningful temporal patterns, e.g., dividing cell clusters during a differentiation process. We train a deep learning model to generate synthetic high-dimensional gene expression profiles corresponding to the simulated pattern at each time point, and apply the different dimension reduction approaches on the high-dimensional time-series data to compare how well they reflect the underlying temporal pattern introduced in, e.g., t-SNE space.

We thus characterize the different perspective of each technique on a specific temporal pattern with respect to the underlying representation in which the pattern was introduced and to the pattern itself. The results illustrate how the choice of the dimension reduction approach can dramatically alter, i.e, distort, temporal structure. To alleviate such problems, we provide directions for designing dimension reduction techniques that explicitly respect temporal structure.

Monday, 04/Sept/2023 12:20pm - 12:40pm
ID: 169 / **S7: 5**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), Statistical hypothesis testing (covariate adjustment, nonparametric methods, multiple comparisons, …), Preclinical drug development, safety and toxicology
Keywords: efficacy-toxicity, equivalence test, generalized joint regression, mixed outcome, similarity test

Testing for similarity of multivariate mixed outcomes with application to efficacy-toxicity responses

Niklas Hagemann[1], Giampiero Marra[2], Frank Bretz[3], Kathrin Möllenhoff[1]
[1]Mathematical Institute, Heinrich-Heine-University Düsseldorf, Germany; [2]Department of Statistical Science, University College London, United Kingdom; [3]Statistical Methodology, Novartis Pharma AG, Basel, Switzerland; niklas.hagemann@hhu.de

A common problem in clinical trials is to test whether an effect of an explanatory variable on the response, e.g. the effect of the dose of a compound on efficacy, is similar between two groups. In this context, similarity is defined as equivalence up to a pre-specified threshold specifying the accepted deviation between the groups. Such question is usually assessed by testing whether the (marginal) effects of the explanatory variable on the response are similar, based on, for example, confidence intervals for differences, or, to mention another example, the distance between two parametric models. These approaches typically assume a univariate continuous or binary outcome variable. An approach for associated bivariate binary response variables, based on the Gumbel model, has been recently introduced (Möllenhoff et al., 2021).

In this talk, we propose a flexible extension of such methodology that builds on a *generalized joint regression* framework with Gaussian copula. Compared to existing approaches, this allows for various scales of the outcome variables (e.g. continuous, binary, categorical, ordinal) including mixed outcomes as well as responses with more than two dimensions. We demonstrate the validity of our approach by means of a simulation study. An efficacy-toxicity case study demonstrates the practical relevance of the approach.

Reference:

Möllenhoff, K., Dette, H., and Bretz, F. (2021). Testing for similarity of binary efficacy–toxicity responses. *Biostatistics*, 23(3), 949–966.

Monday, 04/Sept/2023 11:50am - 12:15pm
ID: 465 / S4: 3
Presentation Submissions - Featured Session
Featured Sessions: Biometrical Journal Showcase - Editor's Selection
Keywords: clinical trial, missing data, k-mean clustering, oversampling, recurrent neural network

Missing data imputation in clinical trials using recurrent neural network facilitated by clustering and oversampling

Halimu Haliduola[1], Frank Bretz[2,3], Ulrich Mansmann[4]
[1]Alvotech Germany GmbH, Germany; [2]Novartis Pharma AG, Basel, Switzerland; [3]Section for Medical Statistics, Medical University of Vienna, Vienna, Austria; [4]Institute for Medical Information Processing, Biometry and Epidemiology (IBE), LMU Munich, Munich, Germany; Halimuniyazi.Haliduola@alvotech.com

In clinical practice, the composition of missing data may be complex, for example, a mixture of missing at random (MAR) and missing not at random (MNAR) assumptions. Many methods under the assumption of MAR are available. Under the assumption of MNAR, likelihood-based methods require specification of the joint distribution of the data, and the missingness mechanism has been introduced as sensitivity analysis. These classic models heavily rely on the underlying assumption, and, in many realistic scenarios, they can produce unreliable estimates. In this paper, we develop a machine learning based missing data prediction framework with the aim of handling more realistic missing data scenarios. We use an imbalanced learning technique (i.e., oversampling of minority class) to handle the MNAR data. To implement oversampling in longitudinal continuous variable, we first perform clustering via k-mean trajectories. And use the recurrent neural network (RNN) to model the longitudinal data. Further, we apply bootstrap aggregating to improve the accuracy of prediction and also to consider the uncertainty of a single prediction. We evaluate the proposed method using simulated data. The prediction result is evaluated at the individual patient level and the overall population level. We demonstrate the powerful predictive capability of RNN for longitudinal data and its flexibility for nonlinearmodeling. Overall, the proposed method provides an accurate individual prediction for both MAR and MNAR data and reduce the bias of missing data in treatment effect estimation when compared to standard methods and classic models. Finally, we implement the proposed method in a real dataset from an antidepressant clinical trial. In summary, this paper offers an opportunity to encourage the integration of machine learning strategies for handling ofmissing data in the analysis of randomized clinical trials.

Tuesday, 05/Sept/2023 2:40pm - 3:00pm
ID: 190 / **S33: 2**
Presentation Submissions - Invited Session
Invited Sessions: Endpoints in clinical trials and medical product development: Multiple endpoints, composite endpoints, and biomarkers and surrogate endpoints
Keywords: Benefit:Risk Evaluation, Composite Endpoint, Wilcoxon-Mann Whitney Statistic

Design and Analyis of Desirability of Outcome Ranking in Clinical Trials

Toshimitsu Hamasaki, **Scott Evans**
George Washington University Biostatistics Center, United States of America; thamasaki@gwu.edu

Typical analyses of clinical trials involve intervention comparisons for each efficacy and safety outcome. Outcome-specific effects are estimated and marginal effects are potentially combined in benefit:risk analyses. It is widely believed that such analyses provide comprehensive information regarding the intervention effects on patients. However such approaches do not incorporate associations between outcomes of interest, suffer from competing risk challenges when interpreting outcome-specific results, do not recognize the cumulative nature of multiple outcomes on individual patients, and since efficacy and safety analyses are often conducted using different analysis populations, the population to which such benefit:risk analyses apply, is unclear.

The Desirability of Outcome Ranking (DOOR) methodology is a paradigm that resolves these challenges. The DOOR methodology allows us to more effectively evaluate and select treatment strategies by providing a more informative way to compare the patient-centric benefits and risks of intervention alternatives. In this talk, we discuss statistical methods for design, analysis and monitoring of clinical trials with DOOR.

Monday, 04/Sept/2023 4:10pm - 4:30pm
ID: 141 / **S17: 1**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Prognostic and predictive biomarkers in personalized medicine
Keywords: Principles for subgroup identification; Cut-points; Exact confidence bands; Alzheimer's Disease; Multiple comparisons

Confident and Logical Selection of the Cut-point of a Biomarker for Patient Targeting

Yang Han
Department of Mathematics, University of Manchester, UK; yang.han@manchester.ac.uk

Confidently choosing a cut-point for a continuously valued biomarker to target patients is challenging, because there are two levels of multiplicity: the multiplicity of efficacy in the marker-positive subgroup and in the marker-negative subgroup at each cut-point, and the further multiplicity of searching through infinitely many cut-points. Currently available methods do not strongly control familywise type I error rate (FWER) across both levels of multiplicity. I will present a method that does. Taking a confidence band approach, our method in fact sets forth four principles that we believe every confident biomarker cut-point selection method should strive to adhere to.

For diseases with continuous outcome such as Type II Diabetes and Alzheimer's Disease, our method provides exact simultaneous confidence intervals for efficacy in the marker-positive and marker-negative subgroups, simultaneously for all possible cut-point values. I will demonstrate an interactive app for it.

Tuesday, 05/Sept/2023 12:00pm - 12:20pm
ID: 144 / **S26: 4**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical hypothesis testing (covariate adjustment, nonparametric methods, multiple comparisons, …)
Keywords: confidence sets, linear regression, pointwise tolerance intervals, simultaneous tolerance intervals, statistical calibration

Statistical calibration for infinite many future values in linear regression: simultaneous or pointwise tolerance intervals or what else?

Yang Han[1], Yujia Sun[1], Lingjiao Wang[1], Wei Liu[2], Frank Bretz[3]
[1]Department of Mathematics, University of Manchester, UK; [2]School of Mathematical Sciences & Southampton Statistical Sciences Research Institute, University of Southampton, UK; [3]Novartis Pharma AG, Basel, Switzerland;
lingjiao.wang@postgrad.manchester.ac.uk

Statistical calibration using regression is a useful statistical tool with many applications. For confidence sets for x-values associated with infinitely many future y-values, there is a consensus in the statistical literature that the confidence sets constructed should guarantee a key property. While it is well known that the confidence sets based on the simultaneous tolerance intervals (STIs) guarantee this key property conservatively, it is desirable to construct confidence sets that satisfy this property exactly. Also, there is a misconception that the confidence sets based on the pointwise tolerance intervals (PTIs) also guarantee this property. This paper constructs the weighted simultaneous tolerance intervals (WSTIs) so that the confidence sets based on the WSTIs satisfy this property exactly if the future observations have the x-values distributed according to a known specific distribution F(·). Through the lens of the WSTIs, convincing counter examples are also provided to demonstrate that the confidence sets based on the PTIs do not guarantee the key property in general and so should not be used. The WSTIs have been applied to real data examples to show that the WSTIs can produce more accurate calibration intervals than STIs and PTIs.

Monday, 04/Sept/2023 4:50pm - 5:10pm
ID: 330 / **S21: 3**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), Free Contributions

On estimation of use efficiency

Jens Hartung, Danilo Crispim Massuela, Hans-Peter Piepho
University of Hohenheim, Germany; jens.hartung@uni-hohenheim.de

Efficient use of resources is an important aspect of sustainable agriculture systems in future. In agronomy, research is performed to increase water and nutrient use efficiency. There is a range of different definitions of use efficiency that vary in calculation and interpretation in detail. These definitions have in common that use efficiency is defined as the relationship between an investigated input, e.g. water or nutrient, and the output reached by application of the input. The relationship can be calculated in two different ways: The most common approach is to calculate the ratio prior to analysis and submit the ratio to some sort of analysis. The alternative is to perform a regression analysis and estimate the slope in the regression of the output on the input. Under certain conditions, both approaches result in identical estimates of use efficiency, but vary in the estimated standard error and therefore in the results from significant tests. The current work aims to cast light on the similarities and differences between both approaches and to develop a suggestion which approach should be used.

Wednesday, 06/Sept/2023 8:30am - 8:50am
ID: 427 / **S43: 1**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Clinical trial designs (adaptive and platform designs, external controls, …), Estimands and causal inference

Generalizing the intention-to-treat effect of an active control from historical placebo-controlled trials to an active-controlled trial

Qijia He[1], Fei Gao[2], Oliver Dukes[3], Bo Zhang[2]
[1]University of Washington; [2]Fred Hutchinson Cancer Center; [3]Ghent University, Belgium; oliver.dukes@ugent.be

In many clinical settings, an active-controlled trial design (e.g., a non-inferiority or superiority design) is often used to compare an experimental medicine to an active control (e.g., an FDA-approved, standard therapy). One prominent example is a recent phase 3 efficacy trial comparing long-acting cabotegravir, a new HIV pre-exposure prophylaxis (PrEP), to the FDA-approved daily oral tenofovir diphosphate plus emtricitabine (TDF/FTC). One key complication in an active-controlled trial is that the placebo arm is lost and the efficacy of the active control (and hence the experimental drug) compared to the placebo can only be inferred by leveraging other data sources. In this work, we propose a rigorous causal inference framework to infer the intention-to-treat (ITT) effect of the active control using relevant historical placebo-controlled trial data of the active control. We highlight the role of adherence and unmeasured confounding, discuss in detail identification assumptions and two modes of inference (point versus partial identification), propose estimators under identification assumptions permitting point identification, and lay out sensitivity analyses needed to relax identification assumptions. We applied our framework to estimating the intention-to-treat effect of daily oral TDF/FTC versus placebo using data from an active-controlled trial (HPTN 084) and an earlier Phase 3, placebo-controlled trial of daily oral TDF/FTC (Partners PrEP).

ID: 472 / **STRATOS 2: 2**
Presentation Submissions - Featured Session
Featured Sessions: Ten years of the STRengthening Analytical Thinking for Observational Studies (STRATOS) initiative – progress and looking to the future

Ongoing research towards state-of-the-art in variable and functional form selection for statistical models

Georg Heinze[1,4], **Aris Perperoglou**[2], **Willi Sauerbrei**[3]
[1]Center for Medical Data Science, Medical University of Vienna, Vienna, Austria; [2]Predictive Modelling, GSK, Stevenage, United Kingdom; [3]Institute of Medical Biometry and Statistics, Medical Center University of Freiburg, Freiburg, Germany; [4]for TG2; georg.heinze@meduniwien.ac.at

Topic group 2 (TG2) of the STRATOS initiative deals with multivariable model building, in particular with issues in building suitable descriptive regression models. In particular, our TG will provide guidance on strategies for variable selection and on the specification of the functional form of nonlinear effects of continuous covariates.

Medical literature is still full of outdated statistical approaches, because there is a lack of awareness of possible pitfalls of commonly used methods even in very common scenarios. Contrary to that, a large number of new methods is proposed, but most of these methods are stuck in an early phase of development (see Heinze et al, 2023, https://doi.org/10.1002/bimj.202200222). There is insufficient evidence of them being fit for purpose. Often guidance for suitable methods is missing, and scientifically inferior procedures are used to analyse medical data, leading to questionable medical conclusions. These problems affect all STRATOS topics, but here we focus on multivariable modelling.

In our overview paper (Sauerbrei et al, 2020, https://doi.org/10.1186/s41512-020-00074-3) we identified seven areas where evidence should be created by well-designed comparison studies. These areas include: (1) investigating properties of variable selection strategies, (2) comparing spline procedures, (3) analysing variables with a spike at zero, (4) comparing multivariable procedures for variable and function selection, (5) clarifying the role of shrinkage to correct for bias induced by data-driven decisions, (6) evaluating approaches to post-selection inference, and (7) adapting model building strategies to large sample sizes.

In this talk we mention recent activities inside and outside STRATOS to address these issues. Research is ongoing in almost all of these seven areas, but it will need further neutral studies exploring the empirical properties of existing methods in a wider range of problems, and studies that are able to uncover situations where established methods may fail and clarify which assumptions of a method are crucial and which are non-critical. We encourage researchers to perform, reviewers to appreciate, and biostatistical journals to publish such carefully planned method evaluation studies that are indispensable to create the evidence for defining a state-of-the-art in multivariable modelling.

Monday, 04/Sept/2023 11:40am - 12:00pm
ID: 171 / **S1: 3**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Neutral comparison studies in methodological research
Keywords: biostatistics, methodological research, reproducibility

Phases of methodological research in biostatistics - Building the evidence base for new methods

Georg Heinze[1], Anne-Laure Boulesteix[2], Michael Kammer[1], Tim Morris[3], Ian White[3]
[1]Center for Medical Data Science, Medical University of Vienna, Austria; [2]Institute for Medical Information Processing, Biometry and Epidemiology, Ludwig-Maximilians University of Munich, Germany; [3]MRC Clinical Trials Unit, UCL, London, UK;
georg.heinze@meduniwien.ac.at

Although new biostatistical methods are published at a very high rate, many of these developments are not trustworthy enough to be adopted by the scientific community. We propose a framework to think about how a piece of methodological work contributes to the evidence base for a method. Similar to the well-known phases of clinical research in drug development, we propose to define four phases of methodological research. These four phases cover (I) proposing a new methodological idea while providing, for example, logical reasoning or proofs, (II) providing empirical evidence, first in a narrow target setting, then (III) in an extended range of settings and for various outcomes, accompanied by appropriate application examples, and (IV) investigations that establish a method as sufficiently well-understood to know when it is preferred over others and when it is not; that is, its pitfalls. We suggest basic definitions of the four phases to provoke thought and discussion rather than devising an unambiguous classification of studies into phases. Too many methodological developments finish before phase III/IV, but we give two examples with references. Our concept rebalances the emphasis to studies in phases III and IV, that is, carefully planned method comparison studies and studies that explore the empirical properties of existing methods in a wider range of problems. All authors of this paper are members of the international STRATOS Initiative (STRengthening Analytical Thinking for Observational Studies). The proposed framework aims at refining the notion of evidence in methodological research that is central to STRATOS' efforts.

Monday, 04/Sept/2023 3:00pm - 3:20pm
ID: 175 / **S14: 4**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Estimands and causal inference
Keywords: Principal stratum, estimand, missing data, ADA

Novel weighted approach for estimating effects in principal strata with missing data in randomized clinical trials

Dominik Heinzmann[1], **Shengchun Kong**[2], **Sabine Lauer**[3], **Lu Tian**[4]
[1]F. Hoffmann-La Roche Ltd.; [2]AbbVie; [3]Dr. Lauer Research; [4]Stanford University; dominik.heinzmann@roche.com

Many clinical variables such as biomarkers can change in a patient after initiation of a therapeutic treatment and can help identify patients who benefit most from the treatment. In a randomized controlled trial (RCT), often such post-randomization variables are only measured in the experimental treatment arm, and may contain a significant amount of missing data. Such a situation can be addressed by a principal stratum estimand strategy (ICH E9 addendum).

We present a novel weighted imputation regression approach (WRI) for this setting with missing data. The methodology and assumption required to yield valid causal inference are discussed. The good performance of WRI is demonstrated via a simulation study and its application to a large number of clinical trials to investigate treatment effect differences among patients developping or not anti-drug antibodies (ADA; the intercurrent event) to a therapeutic treatment.

Monday, 04/Sept/2023 2:20pm - 2:40pm
ID: 388 / **S14: 2**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Free Contributions
Keywords: Replicability, overall Type-I error rate, project power, drug approval, evidence

Beyond the two-trials rule

Leonhard Held
University of Zurich, Switzerland; leonhard.held@uzh.ch

The two-trials rule requires "at least two adequate and well-controlled studies, each convincing on its own, to establish effectiveness". This is usually employed by requiring two significant pivotal trials and is the standard requirement by regulators before new drugs are approved. However, drug applications are often based on more than two trials and some alternatives and generalizations have been recently proposed to properly deal with this case, among them the harmonic mean chi-squared test (Held, 2020, doi: 10.1111/rssc.12410) and the 2-of-3 rule (Rosenkranz, 2022, doi: 10.1007/s43441-022-00471-4). Both can be even extended to more than 3 pivotal trials. I will compare the different approaches in terms of their statistical properties, with a focus on Type-I error rate, project power and expected sample sizes. Type-I error rates will be considered either under the intersection null hypothesis, where all studies are assumed to have a null effect, or the union null hypothesis, where only some of the studies have a null effect. The applicability of the different methods to combine evidence from clinical trials in conditional or accelerated approval will also be discussed.

Monday, 04/Sept/2023 3:00pm - 3:20pm
ID: 429 / S10: 4
Presentation Submissions - Regular Session (Default)
Topic of Submission: Personalized health care, Real world data and evidence
Keywords: Chronic opioid use, clinical prediction model, Brier score, real-world evidence

Predictors for chronic opioid use – real world evidence using insurance claims data from Switzerland

Ulrike Held[1], Tom Forzy[2], Andri Signorell[3], Manja Deforth[1], Jakob M. Burgstaller[4], Maria M. Wertli[5]
[1]Department of Biostatistics, University of Zurich, Switzerland; [2]Department of Global Health and Population, Harvard T.H. Chan School of Public Health, Boston, MA, USA; [3]Department of Health Sciences, Helsana, Dübendorf, Switzerland; [4]Institute of Primary Care, University and University Hospital Zurich, Switzerland; [5]Department of Internal Medicine, Cantonal Hospital Baden, and Department of General Internal Medicine, University Hospital Bern, Switzerland; ulrike.held@uzh.ch

A public health crisis has been observed in the United States resulting from permissive opioid regulation [1]. In Europe, the monitoring system indicates an increased use of opioids and also an increase in opioid addiction from prescribed opioids [2]. Long-term opioid use may result in opioid dependency and opioid-related adverse events, with considerable impact on personal, societal, and economic cost. We aimed to develop and validate a prognostic prediction model for the outcome chronic opioid use, with the intention to identify patients at high risk for chronic opioid use early.

All consecutive adult patients of Helsana insurance company, one of the largest insurance companies in Switzerland, with at least one opioid claim between 2013 and 2018 were included in this study, and morphine equivalent doses (MED) were calculated for all prescriptions [3]. Predictor domains covered socioeconomic variables, disease specific risk factors, prescriber variables, and the initial opioid dose. In an internal-external development and validation approach, using traditional statistical methods and machine learning, a prediction model for the outcome chronic opioid use, defined as an episode duration of > 90 days with ten or more claims, or an episode duration of > 120 days independent of number of claims, was derived. Model performance was assessed with the scaled Brier score to assess overall accuracy, with discrimination using the c-statistic, and calibration plots. Results of our study were reported according to TRIPOD guidelines.

In the real-world data base of Helsana insurance company, 418,625 episodes of opioid prescriptions were observed in a population of 266,476 patients. Seventeen percent of the episodes turned into episodes with chronic opioid use. Using variables from all predictor domains resulted in a model performance that was slightly better using traditional statistical models, i.e., logistic regression, with a c-statistic of 0.927 (95% CI 0.924 to 0.931), and a scaled Brier score of 48.2% in the validation set, whereas the corresponding numbers were 0.909 (95% CI 0.905 to 0.913), and scaled Brier score of 38.1% for the machine learning approach using random forests with bootstrap samples. Our study showed evidence that chronic opioid use can be predicted at the initiation of an opioid prescription episode, with high accuracy using data routinely collected at a large health insurance company. Traditional statistical methods resulted in higher discriminative ability and similarly good calibration as compared to machine learning approaches. These findings are relevant for individualized health care, but they need to be validated prospectively. The net-benefit of the proposed prognostic model for clinical practice needs evaluation with an impact study.

References

1. Humphreys, K., et al., *Responding to the opioid crisis in North America and beyond: recommendations of the Stanford-Lancet Commission*. Lancet, 2022. 399(10324): p. 555-604.
2. Seyler, T., et al., Is Europe facing an opioid epidemic: What does European monitoring data tell us? Eur J Pain, 2021. 25(5): p. 1072-1080.
3. Burgstaller, J.M., et al., Increased risk of adverse events in non-cancer patients with chronic and high-dose opioid use-A health insurance claims analysis. PLoS One, 2020. **15**(9): p. e0238285.

Tuesday, 05/Sept/2023 2:20pm - 2:40pm
ID: 195 / **S34: 2**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Meta-Analysis and Systematic Reviews

Implementation of the anchor-based indirect comparison method for equivalence margin derivation in biosimilar development

Claudia Hemmelmann[1], Jessie Wang[2], Rachid El Galta[1]
[1]Hexal AG, Germany; [2]Sandoz Pharmaceutical; claudia.hemmelmann@sandoz.com

The standard approach for equivalence margin (EQM) derivation is to perform a “classical” meta-analysis on direct comparisons data and use the 95% confidence interval of the pooled effect with maintaining a certain factor of the treatment effect compared to placebo. However, the treatment regimens in many indications becomes more complex (e.g., combination treatments) and for most of these clinical study data, direct comparisons are not available. Study data for the comparison of treatment A vs treatment B as well as treatment B vs treatment C are available in some situations.

An anchor-based indirect comparison can be applied to estimate the treatment effect of treatment A vs treatment C. This treatment effect (A vs C) can be estimated by calculating the difference of the two treatment effects and the variance is the sum of both variances. The 95% confidence interval of this estimated treatment effect can then be used to derive the EQM.

However, the assumptions transitivity and consistency need to be fulfilled. Both assumptions cannot be statistically tested, especially when only two studies are available. To support the use of this anchor-based indirect comparison different sensitivity analyses were performed.

In this presentation we show the derivation of the EQM using the anchor-based indirect comparison along with the sensitivity analyses for a planned efficacy trial in the biosimilar setting. This approach was successfully applied and agreed with agencies.

Tuesday, 05/Sept/2023 2:20pm - 2:40pm
ID: 167 / **S35: 2**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), Epidemiology
Keywords: pertussis, serial serological survey data, censoring

Pertussis in Belgium - The challenge of using historical serial serological survey data

Sereina Herzog[1], Steven Abrams[2,3], Amber Litzroth[4], Heidi Theeten[5], Niel Hens[2,3]
[1]Medical University of Graz, Austria; [2]Hasselt University, Belgium; [3]University of Antwerp, Belgium; [4]Sciensano, Belgium; [5]Agentschap Zorg & Gezondheit, Belgium; sereina.herzog@medunigraz.at

Pertussis or whooping cough is a highly contagious vaccine preventable disease. Incidence of pertussis has known a steady decline after the introduction of pertussis vaccination in the first half of the previous century, nevertheless, pertussis incidence increased over the past two decades in many countries. The analysis of serial serological survey data can improve our understanding about the dynamics of pertussis.

However, the development of assays for the detection of IgG antibodies in sera entails that various assays have been used for different survey years and thus the antibody titre measurements are not directly comparable. Comparable sero-epidemiological results would enable exploring statistical and mathematical models to estimate time-varying epidemiological parameters, i.e. 'standardized' assay results are required. The long-term goal is to investigate the consequences of the uncertainty related to the standardization of pertussis toxin IgG antibodies results from three serological surveys conducted in Belgium (2002, 2006, 2013). In the current project, we focus on finding the standardization model for each survey.

In each survey, 150 samples were selected such that the range of the original values for IgG antibodies against pertussis toxin was as best as possible covered. All 450 samples were then tested using a magnetic bead-based multiplex immunoassay (MIA). We explored different models for standardization in a Bayesian framework using log-transformed titer values. We considered several strategies for dealing with the censored data that occur due to the limit of detection in assays. i.e. in the MIA assay (dependent parameter) as well as in the original assays (independent parameter). Models and strategies regarding censoring are contrasted using fitted curves with confidence bands as well as posterior predictions.

The choice of model for standardization depends on the strategy used for censored data, which can lead to substantial differences in predictions. The uncertainty in the standardization of antibody titres needs to be reflected in models aimed at estimating time-varying epidemiological parameters from serial serological survey data.

Thursday, 07/Sept/2023 8:50am - 9:10am
ID: 258 / **S62: 2**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Epidemic short-term forecasting in real time
Keywords: short-term forecasts, epidemiology, COVID-19, spatio-temporal statistics

Improving short term forecasts of COVID-19 incidence with subnational epidemic indicators

Stefan Heyder, **Thomas Hotz**
TU Ilmenau, Germany; stefan.heyder@tu-ilmenau.de

Short term forecasts of the case incidence are a key component in designing public health countermeasures to control an ongoing epidemic. To perform such forecasts one usually estimates quantities related to the growth of cases, e.g. the reproduction number or the growth factor, which are readily available on the national level. However there is considerable heterogeneity in the speed at which cases proliferate within one country, making estimates of these quantities on the subnational level desirable for accurate forecasts. As subnational estimates pose more difficulties due to small incidences and spatial correlation, we combine epidemic modelling with techniques from small area estimation and spatial statistics to facilitate estimation and thus forecasts. The usefulness of our approach is demonstrated by simulation studies and through application to German case data on COVID-19.

Thursday, 07/Sept/2023 11:00am - 11:20am
ID: 347 / S68: 2
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Design of preclinical experiments
Keywords: d-optimality, preclinical experiments, robustness

D-optimality in preclinical dose-response studies

Leonie Hezler[1], Jan Beyersmann[2], Bernd-Wolfgang Igl[1]
[1]Boehringer Ingelheim Pharma GmbH und Co. KG, Germany; [2]Universität Ulm; leonie_theresa.hezler@boehringer-ingelheim.com

A proper experimental design is key for the validity of statistical results and their interpretations. In pharmaceutical industry research and preclinical experiments build the starting point for further developments and corresponding results are usually based on small sample sizes. One important goal is to determine a precise estimation of the dose-response shape of the test item. This includes a minimal variance of the calculated model parameters leading to a minimal sample size to obtain a desired precision which also takes the “Three Rs Principle” (Replacement, Reduction, Refinement) into consideration.

In this talk, d-optimal design settings to improve the efficiency of an experimental design are discussed and various non-clinical applications are presented. A particular focus is on the practicability as well as the robustness of the statistical outcomes based on moderate model misspecifications. Therefore, optimal design settings are compared to balanced designs in terms of the precision of the estimates, and bayes optimal designs are investigated to account for the underlying model uncertainty during the planning stage. In addition, a multivariate approach, e. g. for the analysis of drug combinations, will be investigated.

Thursday, 07/Sept/2023 8:30am - 8:50am
ID: 312 / S58: 1
Presentation Submissions - Regular Session (Default)
Topic of Submission: Clinical trial designs (adaptive and platform designs, external controls, …), COVID pandemic
Keywords: Clinical trial design, Group sequential designs, Updated analyses, Terminology

Statistical and regulatory lessons learned from the pandemic

Benjamin Hofner[1,2], **Elina Asikanius**[3]
[1]Paul-Ehrlich-Institut (PEI), Germany; [2]FAU Erlangen-Nürnberg, Germany; [3]Finnish Medicines Agency (FIMEA), Finland;
benjamin.hofner@pei.de

During the pandemic, approaches to speed up regulatory approval were implemented by the EMA. These included rapid scientific advices and rolling reviews. We will briefly describe these approaches, and share our thoughts and learnings on the operational and scientific aspects of these procedures.

One specific issue we encountered during the rolling review of the first vaccine trials was centered around group sequential designs and early updated analyses after the trials met their primary objectives. The updates occurred within a very short time frame of around two weeks after the confirmatory analyses but were based on a substantially increased database with about twice the number of COVID-19 cases. This led to the question how to best communicate these results to the wider public and how to appropriately communicate the uncertainty of the updated analyses, e.g. via confidence intervals.

According to the ICH E9 guideline a trial is to be analyzed and the primary hypothesis is to be "tested when the trial is complete" (ICH 1998). However, the primary analysis and the end of trial often do not coincide, e.g. in group sequential designs, which allow confirmatory testing at interim analyses (potentially without stopping the trial) and in time to event trials were the natural end of study is almost never achieved as usually not all participants experience an event. Fixed design trials where updated analyses or a long-term follow-up are foreseen are another example. As terminology for designs with multiple analysis time points (both for GSDs and other designs with updated analyses) often differs between studies, we propose a common terminological framework to improve the communication of study designs and results.

The added value of updated analyses is not always that straightforward. While the information increases with a more mature data set, the uncertainty due to unplanned data cutoffs and lack of type 1 error control increases as well. Difficulties in adequately defining and aligning the primary hypothesis test and the benefit-risk assessment arise. A slightly different but related issue is overrunning, which occurs e.g. if an endpoint is not immediately observed and hence further events might accrue after the trial was stopped. These data need to be taken into consideration at the time of decision making (CHMP 2007). Both topics, updated analyses and overrunning, raise issues e.g. regarding type 1 error control, appropriate reporting of the uncertainty of the results, and the impact on decision making. We discuss methodological and regulatory considerations regarding the planning, analysis and reporting of updated analyses or overrunning especially in the light of the regulatory assessments. The key element is an appropriate pre-specification of analysis time points and methodological considerations in the light of the value of the analyses for the overall procedure.

References

1. ICH (1998). ICH E9 – Statistical principles for clinical trials. CPMP/ICH/363/96
2. CHMP (2007). Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design. CHMP/EWP/2459/02
3. Hofner, Asikanius, et al (submitted). Vaccine development during a pandemic: General lessons for clinical trial design.

Thursday, 07/Sept/2023 8:30am - 9:10am
ID: 112 / S61: 1
Presentation Submissions - Invited Session
Invited Sessions: Statistical strategies in toxicology
Keywords: multiplicity, proof of hazard

The joint analysis of multiple sources of multiplicity in the evaluation of regulatory toxicology bioassays

Ludwig A. Hothorn
retired from Leibniz University Hannover, Germany; ludwig@hothorn.de

The principal contradiction in regulatory toxicology is the need for controlling false negative error rate primarily, i.e. a proof of safety, e.g. using simultaneous non-inferiority tests versus control. But in the routine analysis the proof of hazard is used commonly, i.e. tests for differences vs. control, at level alpha.

In various bioassay many tests are performed at elementary level alpha: i) for different doses, ii) for different times, iii) for different endpoints of the same scale or mostly different-scaled, iv) different sex, v) different effect sizes (e.g., additive, multiplicative), vii) different lm/glm/glmm models (e.g. with baseline covariates or not), viii) different dose metameters in the Tukey trend test (arithmetic, ordinal, ari-log) (Tukey, 1985, Schaarschmidt, 2022) , ix) different tuning parameter in the poly-k-test for mortality-adjusted analysis of tumor proportions, ect..

The first extreme approach *'testing at level alpha each'* generate unacceptable false positive error rates. The second extreme approach *'to control global FWER by multiplicity adjustment*' generate unacceptable false positive error rates. Therefore, a compromise is needed.

In this talk the maxT-test, an UIT, is proposed as a compromise using the empirical variance-covariance matrix from multiple marginal models (Pipper, 2012) (notice, not the correlation between data!). I will demonstrate and interpret simultaneous confidence intervals for several tox assays over certain models/conditions/parameters using the CRAN packages *multcomp*, *MCPAN* and *tukeytrendtest.*

The pros and cons of such approaches are discussed for routine analysis and guidance's in regulatory toxicology. Moreover, the above approach argues also to avoid games with p-values (FWER, FDR, g…) but to consider the underlying marginal models of pre-clinical interest and to interpret it accordingly by well-chosen effect sizes.

References

1. C. B. Pipper, C. Ritz, and H. Bisgaard. A versatile method for confirmatory evaluation of the effects of a covariate in multiple models. Journal of the Royal Statistical Society Series C-Applied Statistics, 61:315–326, 2012.
2. J. W. Tukey, J. L. Ciminera, and J. F. Heyse. Testing the statistical certainty of a response to increasing doses of a
    a. drug. Biometrics, 41(1):295–301, 1985.
3. F. Schaarschmidt, C. Ritz, and L. A. Hothorn. The Tukey trend test: Multiplicity adjustment using multiple marginal models. Biometrics. 2022;78:789–797.

Thursday, 07/Sept/2023 11:40am - 12:00pm
ID: 113 / **S67: 4**
Presentation Submissions - Featured Session
Featured Sessions: Industry meets academia: Session in memory of Dieter Hauschke

Identification of minimal effective dose MED (resp. no observed effect concentration NOEC) in unbalanced designs with possible heterogenous variances- in memory of Dieter Hauschke

Ludwig A. Hothorn
retired from Leibniz University Hannover, Germany; ludwig@hothorn.de

No, I won't start with Dieter's most important paper (InterJClinPharm, 1992). I will focus on a rather niche paper with him, entitled '*Identifying the maximum safe dose: a multiple testing approach*' (JBS, 2000) - nevertheless a relevant issue up to now. First, let me focus on MED (better 'minimal significant dose'), frequently used in both clinical dose finding studies and toxicological/pharmacological bioassays. Second, restrict the alternative to monotone order or not. For order restriction, trend tests, such as Williams test are used. However, pooling contrasts may distort the MED identification (Bauer, 1997) and should avoided. Either related closed testing procedures for monotone effect size assumption (Hothorn & Lehmacher 1991) or the unrestricted, simultaneous Dunnett test can be used instead. However, in balanced and particularly unbalanced designs variance heterogeneity can distort the correct MED identification. Occurs the increased variance in the MED-dose or the control, a natural, fair power loss results. But occurs the increase variance in the non-MED dose, e.g. in the highest dose alone, the MED can be incorrectly too high estimated. Three alternatives to the original Dunnett test work well in these situations: i) using the sandwich estimator instead the MQR estimator (Herberich, 2010), ii) Welch-type modified degree of freedoms (Hasler, 2008) or iii) even Bonferroni-Welch-t-tests (for small number of doses only). Their properties is demonstrated by a tiny simulation study. The sandwich modification should not be used for too small sample sizes.

By means of a real data example and the CRAN packages multcomp, sandwich and Simcomp the appropriate analysis will be demonstrated.

The talk ends with the final recommendation '*Use sandwich/Welch-df modifications for real data evaluation instead of the original Dunnett procedure*'.

References

1. P. Bauer, A note on multiple testing procedures in dose finding .Biometric (53) 1125-1128 (1979)
2. C. W. Dunnett. A multiple comparison procedure for comparing several treatments with a control. JASA, 50(272):1096–1121, 1955.
3. M. Hasler and L. A. Hothorn. Multiple contrast tests in the presence of heteroscedasticity. Biometrical Journal, 50(5):793–800, 2008.
4. D. Hauschke, et al. A distribution-free procedure for the statistical analysis of bioequivalence studies. Int. J. Clinical Pharmacology (30) S37-43 (1992)
5. E. Herberich, J. Sikorski, and T. Hothorn. A robust procedure for comparing multiple means under heteroscedasticity in unbalanced designs. PLOS One, 5(3):e9788, March 2010.
6. L.A. Hothorn, L. & Lehmacher, W. A simple testing procedure "Control versus k treatments" for one-sided ordered alternatives, with application in toxicology. Biom.J. (33) 179-182 (1991)

Monday, 04/Sept/2023 2:00pm - 2:20pm
ID: 147 / S13: 1
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), Time-to-Event Analysis
Keywords: conditional mixed models, marginal models, marginal predictive distributions, survival analysis, categorical data analysis

A transformation perspective on marginal and conditional models

Torsten Hothorn, **Luisa Barbanti**
Universität Zürich, Switzerland; torsten.hothorn@uzh.ch

Clustered observations are ubiquitous in controlled and observational studies and arise naturally in multicenter trials or longitudinal surveys. We present a novel model for the analysis of clustered observations where the marginal distributions are described by a linear transformation model and the correlations by a joint multivariate normal distribution. The joint model provides an analytic formula for the marginal distribution. Owing to the richness of transformation models, the techniques are applicable to any type of response variable, including bounded, skewed, binary, ordinal, or survival responses. We discuss the analysis of two clinical trials aiming at the estimation of marginal treatment effects. In the first trial, the pain was repeatedly assessed on a bounded visual analog scale and marginal proportional-odds models are presented. The second trial reported disease-free survival in rectal cancer patients, where the marginal hazard ratio from Weibull and Cox models is of special interest. An implementation is available in the ``tram'' add-on package to the R system and was benchmarked against established models in the literature.

Reference:

DOI: 10.1093/biostatistics/kxac048

Tuesday, 05/Sept/2023 2:00pm - 2:20pm
ID: 260 / **S35: 1**
Presentation Submissions - Regular Session (Default)
Topic of Submission: COVID pandemic, Statistical modelling (regression modelling, prediction models, …), Epidemiology
Keywords: state space model, epidemiology, COVID-19, monitoring, forecast

State space models as a flexible framework for monitoring epidemics

Thomas Hotz, **Stefan Heyder**
TU Ilmenau, Germany; thomas.hotz@tu-ilmenau.de

To monitor epidemics, different indicators such as incidences, hospitalisation, and deaths play an important rôle. The corresponding data form a multivariate time series whose development over time is governed by several effects: the non-linear reproduction equation describing the spread of the disease, delays due to reporting and the patient-dependent progression of the disease, as well as under-reporting. We show that state space models offer a flexible framework to incorporate these effects. They can easily be fitted even to incomplete data either using an extended Kalman filter or following the likelihood-based approach proposed by Durbin & Koopman (2001). Using publicly available data on COVID-19, we demonstrate how the predicted evolution of states can be used both to monitor the epidemic as well as to generate short-term forecasts.

Wednesday, 06/Sept/2023 11:20am - 11:40am
ID: 267 / **S52: 3**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Machine Learning and Data Science, High dimensional data, genetic and x-omics data

Evaluation of network-guided random forest for disease gene discovery

Jianchang Hu, Silke Szymczak
Universität zu Lübeck, Germany; jianchang.hu@uni-luebeck.de

Identification of biomarkers associated with complex diseases can improve patient risk prediction and foster understanding of underlying molecular pathomechanisms. Gene network information is believed to be beneficial for disease module and pathway identification. We investigate the performance of a network-guided random forest (RF) where the network information is summarized into a sampling probability of predictor variables which is further used in the construction of the RF. The identification of important genes is based on standard variable importance measures from RF. In the simulation study, we simulate synthetic RNA-Seq data along with the underlying network structure using the R package SeqNet. Our results suggest that network-guided RF does not provide better disease prediction than the standard RF. In terms of disease gene discovery, when causal genes are randomly distributed within the network, network information only deteriorates the gene selection, but if they form disease module(s), network-guided RF identifies causal genes more accurately. We also find that when disease status is independent from genes in the given network, spurious gene selection results can occur when using network information, especially on hub genes. Two TCGA microarray and RNA-Seq breast cancer datasets with 283 and 284 patients, respectively, along with protein-protein interaction network information from the STRING database are investigated for progesterone receptor (PR) status related gene identification. Both datasets include 193 PR-positive patients. Standard and network-guided RFs can both find out the core genes including *PGR* and *ESR1* on two datasets. In addition, network-guided RF can further identify gene *EGFR* from the ESR-mediated signaling pathway and gene *AR* from the gene expression (transcription) pathway; both pathways are PGR-related. This demonstrates the potential gains in disease module and pathway identification by utilizing network information for complex diseases.

Tuesday, 05/Sept/2023 2:00pm - 2:20pm
ID: 180 / **S34: 1**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Meta-Analysis and Systematic Reviews
Keywords: Network meta analysis; Clinical trial registries; Missing not at random; Propensity score

IPW-based publication bias adjustment in network meta-analysis with clinical trial registries

Ao Huang
University Medical Center Göttingen (UMG), Germany; ao.huang@med.uni-goettingen.de

Network meta-analysis is an extension of the standard pairwise meta- analysis, which entails us to compare multiple treatments simultaneously and effectively by synthesizing studies of various combinations of comparative treatments. Similarly in the standard meta-analysis, its validity may be threatened by publication bias issue. Although sensitivity analysis methods for publication bias based on selection functions have been developed for pairwise meta-analysis, it is not an easy task to extend them to network meta-analysis due to its complexity. Utilizing clinical trial registries, we propose a simple publication bias adjustment method based on the inverse probability weighting method. Our method can easily handle selective publication processes determined by the t-type statistic of the primary outcome in each study. It is more appealing than the existing sensitivity analysis methods based on the Heckman-type selection function since the t-type statistic would be responsible for publication. Specifically, with the external information from clinical trial registries, we propose multiple estimating equations to estimate the selective publication probability (propensity score) of each study. The maximum likelihood estimation method is inversely weighted by the publication probability to correct the bias. Numerical studies revealed that the proposed method successfully eliminates publication bias. In addition, the adjusted P-score is also proposed for ranking different interventions in the presence of selective publication.

Monday, 04/Sept/2023 12:00pm - 12:40pm
ID: 295 / **S2: 2**
Presentation Submissions - Invited Session
Invited Sessions: Anticipated non-proportional hazards in confirmatory RCTs: should we still use a Cox model and log-rank test?
Keywords: non proportional hazards, systematic review, time-to-event

Methods for non proportional hazards in clinical trials: A systematic review

Cynthia Huber, CONFIRMS consortium
University Medical Center Göttingen, Germany; cynthia.huber@med.uni-goettingen.de

For the analysis of time-to-event data well established methods are available given the assumption of proportional hazards (PH) holds, e.g. log-rank test or Cox PH model. Although a wide range of parametric and non-parametric methods for non-proportional hazards (NPH) is available, no consensus on the best approach exists. We conducted a systematic literature search to identify available statistical methods and software appropriate under NPH.

Our literature search identified 907 abstracts. In our systematic review, we included 211 articles, mostly methodological ones.

The articles include effect measures, effect estimation and regression approaches, hypothesis tests, and sample size calculation approaches which are often tailored to specific NPH situations.

ID: 480 / **STRATOS 1: 2**
Presentation Submissions - Featured Session
Featured Sessions: Ten years of the STRengthening Analytical Thinking for Observational Studies (STRATOS) initiative – progress and looking to the future
Keywords: Initial data analysis, statistical analysis plan, reproducibility

Initial data analysis plans are part of research projects

Marianne Huebner[1,4], **Carsten Oliver Schmidt**[2], **Lara Lusa**[3]
[1]Michigan State University, United States of America; [2]University Medicine of Greifswald, Germany; [3]University of Ljubljana, Slovenia; [4]for TG3; huebner@msu.edu

Initial data analysis (IDA) is a systematic approach that aims for transparency and integrity by providing researchers with an analysis-ready data set and reliable information about its properties. It consists of metadata setup, data cleaning, data screening, initial data reporting, refining or updating the statistical analysis plan, and documenting and reporting IDA. Researchers have flexibility to make decisions throughout a research study, but irreproducibility results when these decisions are handled in an ad-hoc manner. DA is not routinely taught to data analysts thus is often conducted without a clear plan and is not well documented. We discuss a check list for developing a priori IDA plans and illustrate this with examples.

Tuesday, 05/Sept/2023 2:00pm - 2:40pm
ID: 203 / **S33: 1**
Presentation Submissions - Invited Session
Invited Sessions: Endpoints in clinical trials and medical product development: Multiple endpoints, composite endpoints, and biomarkers and surrogate endpoints
Keywords: public health, composite, multi-component, reasonably likely, statistical errors

Endpoint Development and Analysis Planning in Clinical Trials

Hsien-Ming James Hung
US Food and Drug Administration, United States of America; Hsienming.Hung@fda.hhs.gov

Development of endpoints, primary and secondary endpoints, in clinical trials involves consideration of many aspects. Some aspects are whether the endpoints can provide adequate assessments of clinical events (e.g., mortality, morbidity), patient symptoms, measures of function (e.g., ability to walk or exercise), whether the trial can be reasonably powered to study the investigational therapy, whether the endpoints can reasonably well provide public health measures. Recently there seems to be an increasing trend toward adding more components to composite endpoints or multi-component endpoints. In addition, more and more attention has been drawn to the so-called “reasonably likely surrogate” that is a biomarker considered to be able to reasonably likely predict clinical outcomes. Given such greater complexities, handling errors associated with statistical inferences becomes increasing unclear. This talk is planned to focus on what statistical errors to control, measures of substantial evidence, whether other aspects (e.g., doses, heterogeneity of diseases) need to play roles in error control.

Wednesday, 06/Sept/2023 8:30am - 8:50am
ID: 442 / **S49: 1**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Clinical trial designs (adaptive and platform designs, external controls, …), Estimands and causal inference, Epidemiology, Time-to-Event Analysis
Keywords: safety topic of special interest, intercurrent event, competing event

Considerations of safety estimands and estimators in pivotal and post-market studies

Rima Izem, Valentine Jehl, Pedro Lopez Romero
Novartis, Switzerland; rima.izem@novartis.com

The estimand framework, outlined in the ICH-E9 addendum, can help research teams guide their choices among several design and analytical methods, to those that result in better aligned estimators to the main questions or estimands of interest in clinical development. Several recommendations and guidelines exist facilitating the use of this framework to refine efficacy questions or achieve better alignment of estimators to the estimand of interest for pivotal trials. In contrast, few such guidelines exist for refining or aligning with safety estimands whether in pivotal or post-market studies. While several safety methods and estimators exist, such as proportion of events or incidence of events in follow-up, what estimand they align to is often poorly understood in the presence of censoring, intercurrent events, or competing risk.

In this presentation, we demonstrate that the estimand framework is broadly applicable to safety questions, especially relating to those safety topics of special interest, across different phases of development. We will motivate the use of the estimand framework in safety in pivotal, and post-market observational studies. In all studies, we will discuss the impact of safety-specific concepts, such as time-at-risk, rare incidence, and risk mitigation, on the estimand thinking and associated analytical strategies. In observational studies, we will also advocate for the complementarity of the causal inference thinking (e.g., use of the target trial framework concept) and the estimand framework. Finally, we will review and critique existing safety estimators in light of these frameworks and concepts and highlight areas where further methodological work is needed.

Monday, 04/Sept/2023 11:00am - 11:20am
ID: 176 / S3: 1
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), Machine Learning and Data Science
Keywords: Calibration, Logistic regression, Machine learning, Probability estimation, Probability machine, Updating

Calibrating machine learning approaches for probability estimation: a comparison

Max Louis Jansen[1], Francisco Miguel Ojeda[2], Alexandre Thiéry[1], Stefan Blankenberg[1], Christian Weimar[3,4], Matthias Schmid[5], Andreas Ziegler[1,2,6]

[1]Cardio-CARE, Medizincampus Davos, Graubünden, Switzerland; [2]University Heart & Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany; [3]BDH-Klinik Elzach, Baden-Württemberg, Germany; [4]Institute for Medical Informatics, Biometry and Epidemiology, University of Duisburg-Essen, North Rhine-Westphalia, Germany; [5]Institute of Medical Biometry, Informatics and Epidemiology, University of Bonn, North Rhine-Westphalia, Germany; [6]School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, South Africa; max.jansen@cardio-care.ch

Statistical prediction models have gained popularity in applied research. One challenge is the transfer of the prediction model to a different population which may be structurally different from the model for which it has been developed. An adaptation to the new population can be achieved by calibrating the model to the characteristics of the target population, for which numerous calibration techniques exist. In view of this diversity, we performed a systematic evaluation of various popular calibration approaches used by the statistical and the machine learning communities for estimating two-class probabilities. In this presentation, we present the results of a comprehensive simulation study and an application to real data. The calibration approaches are compared with respect to their empirical properties and relationships, their ability to generalize precise probability estimates to external populations and their availability in terms of easy-to-use software implementations. Calibration methods that estimated one or two slope parameters in addition to an intercept consistently showed the best results in the simulation studies. Calibration on logit transformed probability estimates generally outperformed calibration methods on non-transformed estimates. In case of structural differences between training and validation data, re-estimation of the entire prediction model should be outweighted against sample size of the validation data. We recommend regression-based calibration approaches using transformed probability estimates, where at least one slope is estimated in addition to an intercept for updating probability estimates in validation studies.

Thursday, 07/Sept/2023 11:20am - 11:40am
ID: 317 / **S68: 3**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Design of preclinical experiments
Keywords: Reproducibility, experimental design, bias, pseudoreplication

Applying 10 simple rules for good research practice in pre-clinical research

Philip Jarvis
Novartis, Switzerland; philip.jarvis@novartis.com

Applying 10 simple rules for good research practice in pre-clinical research

The reproducibility or lack thereof has been identified as a focus area for science with one of the headlines "*More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments* [1]. There has been much discussion of potential solutions including Recommendations to improve study design and analysis in an NIH White Paper [2]. In this presentation the ideas listed in recently by Schwab et al. [3] will be reviewed and their application within pre-clinical research assessed.

References:

1. Baker, M (2016). Is there a reproducibility crisis? *Nature*, **533**; 452-454. https://www.nature.com/news/polopoly_fs/1.19970!/menu/main/topColumns/topLeftColumn/pdf/533452a.pdf
2. Report of the ACD WORKING GROUP ON ENHANCING RIGOR, TRANSPARENCY, AND TRANSLATABILITY IN ANIMAL RESEARCH, June 11th 2021, https://acd.od.nih.gov/documents/presentations/06112021_RR-AR%20Report.pdf
3. Schwab S, Janiaud P, Dayan M, Amrhein V, Panczak R, Palagi PM, et al. (2022) Ten simple rules for good research practice. PLoS Comput Biol 18(6): e1010139. https://doi.org/10.1371/journal.pcbi.1010139

Wednesday, 06/Sept/2023 9:10am - 9:30am
ID: 343 / S49: 3
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), High dimensional data, genetic and x-omics data, Statistical hypothesis testing (covariate adjustment, nonparametric methods, multiple comparisons, …), Preclinical drug development, safety and toxicology
Keywords: adverse events burden score, Tweedie regression, quantile regression, negative binomial

Adverse event burden score as an alternative approach to quantify and compare adverse event burden in clinical trials

Bartosz Jenner[1], Ming Chen[2], Zhini Wang[3], Shu-Fang Hsu Schmitz [1]
[1]Statistics and Decision Sciences, Janssen Pharmaceutical, Allschwil, Switzerland; [2]Statistics and Decision Sciences, Janssen China Research & Development; [3]IQVIA; BJenner@its.jnj.com

Background:

Adverse events (AEs) in clinical trials (CT) are usually tabulated by frequency and severity. Such separate descriptive summaries cannot reflect the overall AE burden of the treatment and do not provide statistical inference on overall between-treatment differences. The aim of this project is to adopt an AE burden score (AEBS) – a versatile quantitative measure (Le-Rademacher et al. 2020) and to evaluate suitable statistical approaches for comparing treatment effect on AE burden.

Materials/Methods:

For illustration, data from two randomized double-blind placebo controlled CTs (Trials 1 and 2) were explored. The main interest is on the AE burden in the first 12 weeks when the dose was uptitrated weekly to identify the maximum tolerated dose for individual patients. The experimental treatment was known to cause some specific AEs.

The AEBS is a continuous, exposure adjusted metric incorporating duration, severity, and frequency of AEs. For patients without AEs, the AEBS is zero. For some AE episodes information of severity or/and duration was incomplete, therefore AE rate (frequency divided by exposure duration) was also explored. AEBS and AE rates were separately derived for treatment-specific and all AEs.

The distribution of AEBS is skewed with excess zero-values, with mean and variance positively correlated. These features restrict choice of applicable statistical models. For AEBS we evaluated: 1) ANOVA model (reference model); 2) ANOVA model using the natural logarithm of the AEBS plus a small value (delta), i.e., ln(AEBS+delta) to allow including patients without AEs; 3) Tweedie regression; 4) quantile regression for median. For AE rates, 5) negative binomial regression model with exposure duration as offset was evaluated.

The ability of a model to discriminate AE burden (AEBS or AE rate) between treatment groups was assessed based on the ratio of treatment effect and its standard error. Where applicable (models 1-3), fit of the models was compared using the Akaike criterion (AIC).

Results/Discussion:

Trial 1: For models 1-5 the estimated discrimination abilities for all/treatment-specific AEs are: 6.8/14.2; 10.5/16.3; 9.3/17.1; 8.2/14.9; 7.4/16.4. Treatment separation is clearly better for treatment-specific AEs. For AEBS, Tweedie regression is the best for treatment-specific AEs and second-best for all AEs.

Trial 2: For models 1-5 the estimated discrimination abilities for all/treatment-specific AEs are: 3.6/5.3; 4.6/5.7; 5.0/6.1; 5.6/4.8; 5.4/6.9. Treatment separation is better for treatment-specific AEs except the quantile regression. For AEBS, the quantile regression is the best for all AEs. For AE rate, the negative binomial model is good for treatment-specific AEs.

For both trials the Tweedie model has the best fit (AIC).

For future CTs, we recommend AEBS as a versatile and sensitive metric to measure AE burden. Based on the results from the two trials above, Tweedie regression yields good treatment discrimination ability and model fit due to its statistical flexibility to handle frequent zero values and the explicit correlation between mean and variance. A simulation study will be conducted to evaluate whether Tweedie regression is superior to other models under different scenarios.

Wednesday, 06/Sept/2023 11:00am - 11:20am
ID: 374 / S54: 2
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), Software Engineering, Free Contributions
Keywords: simulation, software, review, reproducibility, transparency

An overview of R software tools to support simulation studies: towards standardizing coding practices

Michael Kammer[1,2], Lorena Hafermann[3], Georg Heinze[1]
[1]Medical University of Vienna, Center for Medical Data Science, Institute of Clinical Biometrics, Vienna, Austria; [2]Medical University of Vienna, Department of Medicine III, Division of Nephrology and Dialysis, Vienna, Austria; [3]Charité–Universitätsmedizin Berlin, corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin , Institute of Biometry and Clinical Epidemiology, Berlin, Germany; michael.kammer@meduniwien.ac.at

Simulation studies comparing different approaches to a particular research task play an important role in establishing the evidence base for biostatistical methods. However, conducting such studies is not a trivial task in itself. Key issues discussed in the biostatistical literature comprise the replicability of the simulations, transparent and complete reporting and neutrality. Well-designed and easy to use software tools can help addressing these concerns. However, while software packages exist for many different types of simulation tasks, there seems to be little consensus on how to standardize the actual coding of a simulation study. Consequently, authors of publications often develop their own ad-hoc simulation code.

As a step towards the standardization of coding practices and code sharing, we provide an overview of existing software packages in the programming language R to support simulation studies, with a focus on the coding of the data generating mechanism. We found that there are many powerful and general simulation packages available, but only few of them were accompanied by peer-reviewed publications. Most packages adopted approaches that explicitly specify the data using distributional assumptions, in contrast to methods that create variations of an existing dataset e.g. similar to fully conditional specification.

In addition, we developed an R package for data generation intended to be easy to use, thus lowering the barrier to conducting proper comparison studies for newly developed methods. To complement the existing ecosystem a key goal of our work is to build a library of interesting data generating models derived from real-world datasets, which are then directly and easily available to other users. Such a library of presets serves as starting point for comparison studies and facilitates full replicability and data protection, as well as the standardization of simulation setups by sharing configurations, rather than by sharing full datasets.

We demonstrate a selection of the identified packages including our own by example analyses using real-world datasets for which we derived plausible data-generating models for simulations through the different approaches. Simulation studies are very diverse and therefore a single tool is not sufficient to perform all kinds of such studies. Nevertheless, software packages may facilitate the standardization and exchange of code, thereby providing a framework essential to design better simulation studies.

Thursday, 07/Sept/2023 8:50am - 9:10am
ID: 151 / **S63: 2**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Biomarkers and diagnostics, High dimensional data, genetic and x-omics data
Keywords: pathway, biomarker

Pathway analysis for multinomial phenotypes

Md. Kamruzzaman[1], Taesung Park[2]
[1]Seoul National University, Korea, Republic of (South Korea); [2]Seoul National University, Korea, Republic of (South Korea);
tspark@stats.snu.ac.kr

Many statistical methods for pathway analysis have been used to identify novel pathways from biomarkers associated with a certain disease. However, most of these methods are based on single pathway analysis and do not consider multiple pathways simultaneously. Since pathways are highly correlated, multiple pathways analyses suffer from this correlation. Furthermore, they mainly focus on only continuous, counts, and binary phenotypes. In this study, we propose a novel pathway analysis HisCoM-Categ for the multinomial phenotypes such as the obesity level observed as normal, overweight, and obese. HisCoM-Categ takes into account the hierarchical structure of biomarkers and pathways, as well as the correlations among pathways. Through the simulation study, HisCoM-Categ was shown to have higher power compared to the other existing methods. In addition, HisCoM-Categ was applied to the various types of omics data. This application demonstrated that HisCoM-Categ successfully identified the well-known pathways that are associated with multinomial phenotypes.

Wednesday, 06/Sept/2023 12:00pm - 12:20pm
ID: 235 / **S52: 5**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Epidemiology, Machine Learning and Data Science
Keywords: Generative Modeling, Data Synthesis, Machine Learning

Generative modeling of epidemiological data using adversarial random forests

Jan Kapar[1,2], Kathrin Günther[1], David S. Watson[3], Marvin N. Wright[1,2,4]
[1]Leibniz Institute for Prevention Research and Epidemiology - BIPS; [2]University of Bremen; [3]King's College London; [4]University of Copenhagen; kapar@leibniz-bips.de

Generative modeling of epidemiological data using adversarial random forests

Generative modeling holds great potential for epidemiological data as it opens the door for applications like realistic data imputation for missing data, data augmentation for enhancing predictive performance and privacy-preserving data analysis. However, while deep learning algorithms such as variational autoencoders (VAEs) and generative adversarial networks (GANs) have shown ground-breaking results generating realistic synthetic image, audio and text data during the last decade, these methods often struggle to produce high quality synthetic tabular data. Further, deep learning algorithms are notoriously data-hungry and require extensive tuning.

We present the concept of adversarial random forests (ARFs), a method based on unsupervised random forests that shows promising results for tabular data with both continuous and categorical features. Unlike many deep learning methods, ARFs perform well without expensive hyperparameter tuning and often show good results also on comparably small datasets. Training time for ARFs is considerably shorter than for state-of-the-art deep learning models for tabular data.

To evaluate the utility of synthetic data created with ARFs in real world epidemiological applications, we replicate statistical analyses of already published studies based on the German national cohort dataset (NAKO). We demonstrate that ARFs are capable of successfully learning the underlying structures of the data so that the results of descriptive, inferential and predictive tasks performed on ARF-synthesized data are comparable to the results obtained on the original data and excel in comparison with state-of-the-art deep learning models.

Wednesday, 06/Sept/2023 11:00am - 11:20am
ID: 126 / **S55: 2**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Preclinical drug development, safety and toxicology
Keywords: Dose-response modelling, literature review, alert concentrations

Literature review of dose-response analyses in toxicology

Franziska Kappenberg[1], **Jan G. Hengstler**[2], **Kirsten Schorning**[1], **Jörg Rahnenführer**[1]
[1]TU Dortmund University, Germany; [2]IfADo, Germany; kappenberg@statistik.tu-dortmund.de

Dose-response (or concentration-response, time-response) analyses are an integral part of toxicological research. Often, the goal is to find the lowest condition, where a (significant) change of the response in comparison to a negative control can be observed. For this, both observation-based methods (e.g. Dunnett-test, LOEC) or model-based methods (e.g. ED-values, BMD) are used. For viability assays, parametric modelling with sigmoidal models is well-established. In recent methodological research, parametric modelling and calculation of alert doses based on this parametric modelling has been adressed also for gene expression data

In this talk a review of dose-response analyses published in 2021 in three major toxicological journals is presented. Dose-response analyses from published figures were included when at least four concentrations were measured, where the control is also counted.The review was performed in terms of the biological background (the type of assay, the type of exposition), in terms of the design (the number of considered conditions and the actual condition values, as well as the sample sizes) and in terms of the statistical analysis (the display, the analysis goal, the used methods for testing/modelling and the alert dose of interest). The results from this review are presented comprehensively.

Supported by the findings from the review, a comprehensible guidance is provided, which aspects to consider when designing and analysing dose-response analyses in toxicological research.

ID: 485 / **Plenary 1: 1**
Presentation Submissions - Featured Session
Featured Sessions: Keynote
Keywords: Causal inference, Survival analysis, Longitudinal data, Observational data

Causal inference with observational data: A survival guide

Ruth Keogh
London School of Hygiene & Tropical Medicine, United Kingdom; ruth.keogh@lshtm.ac.uk

Causal inference methods for estimating effects of treatments or other interventions on health outcomes have seen rapid and extensive developments in recent years. Tasks for making causal inferences range from estimating average treatment effects, to conditional average treatment effects, to obtaining individualised predictions under different treatment choices. The types of treatment strategies we may be interested in range from giving a one-time treatment, to assigning a treatment to be sustained over a long period, to dynamic strategies that involve initiating a treatment when the patient reaches a certain health state.

This talk will attempt to provide a guide to the challenges we face in investigating causal effects of treatments on time-to-event outcomes using longitudinal observational healthcare data, and how they can be tackled. I will discuss the importance of clear specification of the research question and what we aim to estimate (the estimand), which in the time-to-event context often involves consideration of competing events.

I will give an overview of some of the statistical methods at our disposal for answering questions about the effects of treatments on time-to-event outcomes, which include marginal structural models, inverse probability weighting, censoring-and-weighting, g-formula, and extensions that offer double-robustness. Traditional methods for investigations into effects of exposures or treatments on survival, such as Cox regression, are sometimes stated as not being suitable for use in causal inference. However, I will discuss how these traditional techniques still form a fundamental part of the toolkit for causal inference for time-to-event outcomes.

A running example in the context of type 2 diabetes will be used, making use of an open access data set designed to mimic longitudinal observational data such as from longitudinal cohort studies, patient registries and electronic health records.

Tuesday, 05/Sept/2023 3:20pm - 3:40pm
ID: 221 / S30: 5
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), Prevention and handling of missing data, Machine Learning and Data Science, Time-to-Event Analysis
Keywords: prediction model, survival, multiple imputation, random survival forest

Developing a survival prediction model – a case study

Samuel Kilian[1], Kathrin Burgmaier[2,3], Max Liebau[4], Meinhard Kieser[1]
[1]Heidelberg University, Heidelberg, Germany, Germany; [2]Department of Pediatrics, Faculty of Medicine, University Hospital Cologne and University of Cologne, Cologne, Germany; [3]Faculty of Applied Healthcare Science, Deggendorf Institute of Technology, Deggendorf, Germany; [4]Department of Pediatrics, Center for Family Health, Center for Rare Diseases, and Center for Molecular Medicine, University Hospital Cologne and Faculty of Medicine, University of Cologne, Cologne, Germany; kilian@imbi.uni-heidelberg.de

The development of prediction models requires a careful choice of methods. This is especially true for survival endpoints since the general practice in regression analysis to predict the mean of a distribution may not be suitable. The TRIPOD statement provides a framework for specifying and reporting the process. This includes aspects like outcome, predictors, missing data, model specification, and validation.

In this talk we discuss pros and cons of different ways to handle each aspect and we present the choices we made for developing and validating a prediction model for kidney survival of patients with the Autosomal Recessive Polycystic Kidney Disease.

For example, the type of prediction made for a patient could be a relative risk score, a complete survival distribution, or something in between. Furthermore, the type of model has to be chosen. While the commonly used Cox model is easily applicable, machine learning methods like random survival forests may give better predictions due to their flexibility. If a small set of predictors is desirable, some kind of variable selection has to be performed. The metric to assess model performance has to be chosen carefully considering the prediction objective. When missing values are handled by multiple imputation, technical details like the pooling of Kaplan-Meier curves have to be considered. Model validation should be prespecified and can be done on a separate validation set or within the development process by cross validation. When a dataset is split into development and validation set, representativeness of both sets should be ensured.

Thursday, 07/Sept/2023 9:30am - 9:50am
ID: 379 / **S60: 4**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …)
Keywords: Shrinkage, variable selection, lasso, simulation study

Post-estimation shrinkage in full and selected linear regression models in low-dimensional data revisited

Edwin Kipruto, Willi Sauerbrei
Medical Center University of Freiburg, Germany; edwin.kipruto@uniklinik-freiburg.de

The fit of a regression model to new data is often worse than its fit to the original data due to overfitting. Analysts often employ variable selection techniques when developing a regression model, which can lead to biased estimates. To address overfitting and reduce the bias of estimates induced by variable selection, shrinkage methods have been proposed. Selected variables whose true effects are small are prone to selection bias and can benefit from shrinkage, while variables with large effects generally require little or no shrinkage. Post-estimation shrinkage is a two-step alternative to penalized regression methods that does not rely on optimizing any criteria under specific constraints and can be easily applied to generalized linear models and regression models for survival data. In the context of a full model and aiming to derive a good predictor, global shrinkage was proposed. For selected models it was extended to parameterwise shrinkage (PWSF). Van Houwelingen and Sauerbrei (2013) conducted a simulation study to compare these two approaches with Lasso, but only for few scenarios with moderate to large signal-to-noise (SNR) ratio and low correlation.

Within the framework of a classical linear regression model, we conducted a simulation study with a much broader scope, specifically concerning the amount of information in the data. We assessed whether post-estimation shrinkage methods can improve full and selected models and compared the results with ridge (in full models) and Lasso (in selected models). We also proposed a modified version of PWSF called nonnegative PWSF (NPWSF) to address the weaknesses of PWSF in full models. We investigated prediction errors, bias of estimates, and model sparsity. The results indicate that the performance of methods is influenced by the amount of information in the data, and none of the methods performed best in all scenarios. Post-estimation shrinkage methods can improve the prediction accuracy of both full and selected models and reduce the bias of regression estimates for selected variables.

In full models, PWSF generally performed poorly, while global shrinkage performed similarly to NPWSF in low SNR. However, in moderate to high SNR, NPWSF outperformed global shrinkage. In addition, NPWSF performed better than ridge in low correlation with moderate to high SNR. In selected models, all post-estimation shrinkage methods performed similarly, with global shrinkage being slightly inferior. Lasso outperformed all post-estimation shrinkage methods in low SNR and high correlation but was inferior in low correlation with high SNR.

Our study suggests that, provided the amount of information is not too small, NPWSF is more effective than global shrinkage in improving the prediction performance of both full and selected models. However, in high correlation or very low SNR, penalized methods appear to outperform post-estimation shrinkage methods.

Reference:

van Houwelingen, H. C., and Sauerbrei, W. (2013). Cross-validation, shrinkage and variable selection in linear regression revisited.

Tuesday, 05/Sept/2023 3:20pm - 3:40pm
ID: 455 / S32: 5
Presentation Submissions - Featured Session
Featured Sessions: Young Statisticians Sessions and Panel Discussion
Keywords: survival analysis, machine learning, calibration, systematic review

The best of two worlds? A systematic comparison of time-to-event model implementations between R and Python

Lukas Klein, Gunter Grieser, Antje Jahn
University of Applied Sciences Darmstadt, Germany; lukas.klein@h-da.de

The German organ transplant registry (TxReg) has recently become available, offering a unique opportunity to study the post-transplant survival of organ recipients in a distinct patient population and healthcare system. In recent years an increasing interest in applying machine-learning (ML) methods for predicting post-transplant survival was observed. However, the issue of censoring in survival analysis prediction tasks seems to be often neglected in machine learning applications. Instead, the task is reduced to classification, a simplification that is only rarely seen in regression modeling. A potential reason for this discrepancy might be a lack of implementations and accompanying documentation in the typically applied machine learning ecosystem of Python for survival analysis. In the field of biostatistics R is more prominent. While R survival analysis packages have a decade-long history, the often-used PySurvival was discontinued in 2019, and only recently has work begun on scikit-survival to integrate survival analysis ML methods into the scikit-learn ecosystem. To guide researchers in selecting and applying ML software for survival predictions, we systematically compare the respective R and Python implementations. We start with comparing the availability of different ML implementations for survival analysis between R and Python. Our inspection also includes a comparison of tools for model inspection and investigating prediction performance with respect to discrimination, accuracy and calibration. Finally, we also compare computational speed when applied to large data. All comparisons are performed and illustrated on TxReg data analyzing kidney recipient post-transplant survival. Our findings show that the Python scikit-survival toolkit provides excellent possibilities for a central interface for implementing survival analysis modelling. In particular the integration of deep learning methods for censored data is an advantage. However, R still provides better tools for model inspection, for assessing prediction performance for example by calibration curves, and for scoring frameworks like TRIPOD. In some cases, such as the ranger package for random survival forests, Python and R share the same backend implementation. In our application, ML methods achieved an IPC weighted C-index of up to 0.72 with a gradient-boosted model. Given the increasing demand for ML and deep learning approaches and the deployment of models in commercial settings, the necessity for survival analysis methodology implemented in Python is expected to grow. Overall, our talk highlights the strengths and limitations of both R and Python for survival analysis and provides guidance for researchers to choose the appropriate toolkit for their analysis needs.

Monday, 04/Sept/2023 11:00am - 12:00pm
ID: 266 / **S2: 1**
Presentation Submissions - Invited Session
Invited Sessions: Anticipated non-proportional hazards in confirmatory RCTs: should we still use a Cox model and log-rank test?
Keywords: time to event, nonproportional hazards, simulation

A simulation-based comparison of statistical methods for time-to-event data analysis under non-proportional hazards

Florian Klinglmueller, CONFIRMS Consortium
AGES - Austrian Agency for Health and Food Safety, Austria; florian.klinglmueller@ages.at

This study presents the results of a comprehensive simulation study that evaluates the performance of statistical methods for time-to-event analysis under non-proportional hazards (NPH). The study covers a wide range of plausible distributional assumptions and typical design options, and compares the operating characteristics of selected statistical methods for testing and estimation in clinical trials with time-to-event endpoints under NPH. The selection of methods and simulation scenarios is based on a systematic review of the scientific literature to identify available options for methods for testing and estimation under NPH and a review of past marketing authorization procedures reporting results from 18 distinct trials. The study covers four broad scenario classes - crossing hazards, delayed onset of treatment effect, progression, and differential treatment effects in subgroups. The findings have important regulatory implications for clinical trials that are pivotal for drug development and benefit-risk assessment under NPH. Additionally, an open-source software package was developed to facilitate simulation of time-to-event data.

Monday, 04/Sept/2023 11:20am - 11:40am
ID: 293 / **S3: 2**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), High dimensional data, genetic and x-omics data
Keywords: Polygenic risk scores, variable selection, prediction, high-dimensional data

Advanced statistical modelling for polygenic risk scores by incorporating alternative loss functions

Hannah Klinkhammer[1,2], Christian Staerk[1], Carlo Maj[3], Peter Krawitz[2], Andreas Mayr[1]
[1]Institute for Medical Biometry, Informatics and Epidemiology, University Hospital Bonn, Germany; [2]Institute for Genomic Statistics and Bioinformatics, University Hospital Bonn, Germany; [3]Center for Human Genetics, Philipps University Marburg, Germany; hannah.klinkhammer@uni-bonn.de

In clinical genetics, it is of interest to predict a trait or phenotype based on the patient's genetic information. Polygenic risk scores (PRS) are based on common genetic variants with low to medium effect sizes and aim to capture this genetic predisposition. As genotype data are high-dimensional in nature, from a technical perspective it is crucial to develop algorithms that can be applied on large-scale data (large n *and* large p). A wide range of PRS methods focus on summary statistics from genome-wide association studies (GWAS) based on univariate effect estimates and combine them to a single score (e.g. PRScs, LDpred2, lassosum). More recently, methods have been developed that can be applied directly on individual-level genotype data to model the variants' effects simultaneously (e.g. BayesR, snpnet). In this context, we introduced snpboost, a framework that applies statistical boosting on individual-level genotype data to estimate PRS directly via multivariable regression models. By iteratively working on batches of variants, snpboost can deal with large-scale cohort data, e.g. from the UK Biobank.

As the technical obstacles are therefore solved, the methodological scope can be now broadened – focusing on the objectives that are really key for the clinical application of PRS. Similar to many other methods, so far, also snpboost has focused solely on quantitative and binary traits based on common loss functions such as the squared error and logistic loss functions. Exploiting the modular structure of statistical boosting, we now incorporated alternatives. As the loss function defines the type of regression problem that is optimized, we effectively extended the snpboost framework to further data situations such as time-to-event and count data. Furthermore, alternative loss functions allow us to focus not only on the mean of the conditional distribution but also on other aspects that may more helpful in the risk stratification of individual patients. In particular, we illustrate two main applications:

First, for time-to-event data types, it is of interest to stratify the lifetime risk with respect to the genetic predisposition, e.g. to implement earlier preventive examinations. In the field of PRS modelling it is common practice to derive a PRS for the binary response of the occurrence of a disease and, in a second step, incorporate this PRS in a Cox proportional hazard model to stratify lifetime risk. In contrast to this approach, we specifically model time-to-event data by using appropriate loss functions (i.e. weighted L_2-boosting or Cox proportional hazard models) and show that optimizing the PRS directly with respect to the aim of predicting the course of the disease is favorable for time-to-event data. Secondly, we include a loss function to fit quantile regression based on the snpboost framework. While most commonly-used methods only provide point estimates for a trait, quantile regression enables us to construct individual prediction intervals quantifying the uncertainty of the prediction for a single patient. Furthermore, quantile regression includes median regression as a special case. Median regression which is a robust alternative to classical mean regression and might be more suitable for traits with outlier measurements.

Tuesday, 05/Sept/2023 2:40pm - 3:00pm
ID: 276 / S31: 3
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Interpretable machine learning in biostatistics: Methods, applications and perspectives
Keywords: Machine Learning, Neural Networks, IML, Feature Attribution

Interpreting Neural Networks: A Biostatistical Perspective

Niklas Koenen[1,2], **Marvin N. Wright**[1,2,3]
[1]Leibniz Institute for Prevention Research and Epidemiology - BIPS; [2]University of Bremen; [3]University of Copenhagen;
koenen@leibniz-bips.de

Throughout the past decade, neural networks have unleashed a tremendous surge of attention and infiltrated almost all conceivable domains of science, medicine, and public life. However, for sensitive applications – besides predictive performance – an understanding of the black box decision process is essential to assess its reliability, gain insights, or extract knowledge from the data. Due to the complexity of deep neural networks, well-established statistical methods and model-agnostic approaches are challenging and often too computationally intensive to apply.

Driven by this lack of neural network-specific interpretability methods, many techniques have been proposed to fill this gap. In this context, so-called feature attribution methods have been developed to reveal variable-wise insights and effects captured in the black box model. Even though these approaches primarily focused on image data and their explanatory quality was frequently assessed by visual impressions, many of these techniques generalize to tabular data, e.g., used in biometric areas. However, the critical question of this generalization is whether and in which situations, feature attribution methods provide reliable explanations and trustworthy data insights. From a biostatistical perspective, we review these methods and challenge them in a simulation study with a known data-generating process and discuss potential pitfalls that may arise in real-world applications. Equipped with the revealed theoretical guidelines, we examine how these methods perform on actual biomedical data and whether they reflect results already found by more established approaches.

Monday, 04/Sept/2023 5:30pm - 5:50pm
ID: 227 / **S16: 4**
Presentation Submissions - Invited Session
Invited Sessions: Causal discovery with a view to the life sciences
Keywords: Bayesian inference, causation, computation, DAG

Bayesian inference of causal graphs: where we are and where we should go

Mikko Koivisto
University of Helsinki, Finland; mikko.koivisto@helsinki.fi

Causal discovery aims at inferring cause–effect relationships between variables from observational data. Recently, there has been notable progress in Bayesian inference of causal graphs, which holds the promise of fully quantifying the uncertainty over competitive causal hypotheses. In this talk, we will highlight the power of the Bayesian paradigm for modeling and inference when the models of interest are only partially identifiable from data. On the other hand, we will also critically examine the assumptions currently needed for statistically and computationally efficient Bayesian inference, including the assumption that we have measured all the common causes of the measured variables. Finally, we will ask how causal discovery in life sciences differs from that in physical sciences or in social sciences.

Tuesday, 05/Sept/2023 4:30pm - 4:50pm
ID: 224 / S42: 2
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), Machine Learning and Data Science
Keywords: Data privacy, Perturbation analysis, Statistical disclosure control, Synthetic data, Validation studies

A simple-to-use R package for mimicking study data by simulations

Giorgos Koliopanos[1], Francisco M. Ojeda[2], Andreas Ziegler[1,2,3]
[1]Cardio-CARE, Medizincampus Davos, Switzerland; [2]Department of Cardiology, University Heart & Vascular Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany; [3]School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Pietermaritzburg, South Africa; george.koliopanos@cardio-care.ch

Background: Data protection policies might prohibit the transfer of existing study data to interested research groups. To overcome legal restrictions, simulated data can be transferred which mimic the structure but are different from the existing study data.

Objectives: The aim of this work is to introduce the simple-to-use R package modgo which may be used for simulating data from existing study data for continuous, ordinal categorical, and dichotomous variables.

Methods: The core is to combine rank inverse normal transformation with the calculation of a correlation matrix for all variables. Data can then be simulated from a multivariate normal and transferred back to the original scale of the variables. Unique features of modgo are that it allows to change the correlation between variables, to perform perturbation analysis, to handle multicenter data, and to change inclusion/exclusion criteria by selecting specific values of one or a set of variables. Simulation studies on real data demonstrate the validity and flexibility of modgo.

Results: modgo mimicked the structure of the original study data. Results of modgo were similar with those from two other existing packages in standard simulation scenarios. modgo's flexibility was demonstrated on several expansions.

Conclusions: The R package modgo is useful when existing study data may not be shared. Its perturbation expansion permits to simulate truly anonymized subjects. The expansion to multicenter studies can be used for validating prediction models. Additional expansions can support the unraveling of associations even in large study data and can be useful in power calculations.

Tuesday, 05/Sept/2023 2:40pm - 3:00pm
ID: 283 / **S29: 3**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Generalized pairwise comparisons
Keywords: GPC, Rank procedures, small samples, Resampling

Applications of generalized pairwise comparisons and rank-based procedures in small samples: Bootstrap and permutation tests

Frank Konietschke
Charite Berlin, Germany; Frank.Konietschke@charite.de

Small samples occur in a variety of different areas and especially in pre-clinical research and translational trials. Most statistical procedures rely on asymptotic results and are thus applicable in large samples only. In case of small samples, they tend to not control the type-I error rate and over-reject the null hypothesis. In addition, postulating a certain data distributional model (e.g. normally distributed data) is often misleading. Nonparametric rank-based methods on the contrary do not rely on any distributional assumption and are thus applicable in these scenarios. However, which effects underlie such methods and which hypotheses are actually tested? In this talk, we discuss such methods in detail, focus on the use of Bootstrap, and (studentized) permutation tests as approximate solutions for small samples. Extensive simulation illustrate the applicability of the methods in (very) small samples. Real data sets illustrate the applications.

Tuesday, 05/Sept/2023 3:00pm - 3:20pm
ID: 353 / **S32: 4**
Presentation Submissions - Featured Session
Featured Sessions: Young Statisticians Sessions and Panel Discussion
Keywords: random forests, variable selection, hyperparameters

Analysis of the Effect of Hyperparameters on Variable Selection in Random Forests

Lea L. Kronziel, Césaire J. K. Fouodo, Inke R. König, Silke Szymczak
Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck, Germany; l.kronziel@uni-luebeck.de

Background: Random forests (RF) are a well-known and efficient method to predict a class or a regression value. RFs are particularly useful for high-dimensional datasets and thus also for analyses of genetic data. In addition, RFs provide the possibility to estimate the importance of each variable. Variable selection algorithms such as Vita and Boruta use these importance scores to statistically test each variable for its relevance for prediction. By selecting variables that are important for the prediction of a disease, insight into the genetic background as well as the biological effects of a disease can be gained.

Motivation: Several studies have shown that the hyperparameters of RFs have an impact on the prediction accuracy. Moreover, the calculated variable importances can be influenced by the choice of hyperparameters. However, it is unclear whether this applies analogously to variable selection procedures. Therefore, the aim of this study was to investigate the influence of different hyperparameters on variable selection by comprehensive simulation studies. Based on the results, recommendations will be given on how to select hyperparameters under specific conditions.

Methods: A large number of simulations were performed within three simulation studies. The focus was on gene expression data since these are usually high dimensional and it is a realistic application area. Initially, we focused on the correlated structure of these data, which were simulated in combination with a continuous target variable. For the simulations with binary target variables, real expression data was used to simulate the structure of the data. The hyperparameters considered were the number of decision trees, the number of split candidates, also known as mtry, and the minimal node size. For variable selection, the Vita and Boruta methods were used, since they have been recommended as those with the highest power. Various performance measures such as false discovery rate (FDR) and sensitivity were used to evaluate the impact of the hyperparameters on Boruta and Vita.

Results: As the number of trees increases, not only the sensitivity increases but also the FDR. Moreover, it turned out that the optimal number of split variables strongly depends on the proportion of associated variables as well as their effect sizes. Furthermore, the minimal node size does not show any relevant effects on variable selection.

Conclusion: The study demonstrates that the previous default values may lead to suboptimal results if the focus is not on prediction performance but on variable selection. As with classification or regression, it is recommended to significantly increase the number of trees for high sensitivity. Strategies must be developed to select optimal values for the number of split candidates for a specific data set.

Thursday, 07/Sept/2023 9:10am - 9:30am
ID: 365 / **S62: 3**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Epidemic short-term forecasting in real time
Keywords: Agent based epidemic models, random graphs, stochastic processes, microsimulations

Predicting the unpredictable: the MOCOS large scale agent based epidemic model

Tyll Krueger, Marcin Bodych, Tomasz Ozanski, Radoslaw Idzikowski
Wroclaw University of Science and Technology, Poland; tyll.krueger@googlemail.com

The MOCOS agent based model is an advanced continuous time epidemic model which is in use for policy recommendation for the polish ministry of health since the beginning of the COVID-19 pandemic. Although the main focus of the model was and is risk analysis and forecasting for the COVID -19 pandemic, it can easily be adapted to the class of respiratory diseases. A special focus of the MOCOS model is the detailed representation of the effect of non pharmaceutical interventions like contact tracing, smartphone based tracing , regional lockdowns and school testing. We present a short overview about the main algorithmic features of the MOCOS model and review the forecast performance of the model based on the submissions to the European forecast hub. We also discuss the Poland specific difficulties of model based policy recommendations. We close the presentation with some challenges of short and medium forecasting for agent based models .

Tuesday, 05/Sept/2023 3:00pm - 3:20pm
ID: 284 / S31: 4
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Interpretable machine learning in biostatistics: Methods, applications and perspectives
Keywords: survival analysis, explainable artificial intelligence, interpretable machine learning, interpretability

Explainability of machine learning models for survival analysis: current state and challenges

Mateusz Krzyziński[1], **Przemysław Biecek**[1,2]
[1]MI2.AI, Warsaw University of Technology, Poland; [2]MI2.AI, University of Warsaw, Poland; mateusz.krzyzinski.stud@pw.edu.pl

The prognostic capabilities of machine learning models for survival analysis match or even surpass those of classical statistical learning approaches like Cox Proportional Hazard models. However, the widespread use of ML models is hindered by their high complexity and lack of interpretability. They are considered black-box, i.e., it is not possible to know directly what influences their prediction internally. Especially in biostatistics, where time-to-event analysis constitutes a fundamental task, there is a demand for techniques that enable the analysis and explanation of machine learning survival models – so-called explainable artificial intelligence (XAI) or interpretable machine learning (IML) methods.

One of the first such techniques was the SurvLIME method [Kovalev et al., 2020], which aims to approximate a complex model using a surrogate model – a well-established Cox model whose coefficients are interpretable and constitute an explanation. It was then expanded with further refinements using the same intuition but with a different optimization context. However, another notable approach is to use time-dependent survival explainable machine learning methods like SurvSHAP(t) [Krzyziński et al., 2023], which decomposes the model's prediction into the effects of individual covariates. Such techniques scrutinize the models' behavior across varying time horizons by analyzing survival or cumulative hazard functions. This time aspect is particularly significant since many ML models do not assume proportional hazards. Thus, analyzing such explanations allows for checking how elastic models reason when this assumption is violated. It also helps uncover uncommon effects of covariates, such as an effect that changes from positive to negative over time. In this talk, we will describe both these methods and demonstrate how the SurvSHAP(t) results can be aggregated to draw conclusions about a model of interest in a global context. This can be achieved by applying functional data analysis concepts and statistical tests to assess the significance of covariate effects in machine learning models.

However, while initial survival XAI methods have shown promising results, several challenges remain to overcome. One such challenge is the computational complexity of analyzing data at multiple time points, which can be both difficult and time-consuming. Furthermore, current methods often rely on analyzing simple right-censored data without accounting for competing risks. To advance the field, it is also essential to establish a stronger conceptual connection between explanations and the underlying biological processes that influence the event of interest. While IML methods can provide explanations for model predictions, they do not directly explain the biological mechanisms that drive the data generating process. We believe that this talk will spark discussion about addressing these challenges and gaps in understanding. It will help improve the effectiveness of XAI methods for survival analysis and ultimately enable more accurate predictions in this crucial field.

Tuesday, 05/Sept/2023 11:40pm - 12:00pm
ID: 439 / **S24: 4**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Machine Learning and Data Science
Keywords: Two-stage and one-stage detectors, precision, pollen taxa recognition

One-stage and two-stage detectors comparison in the task of pollen grains recognition

Elżbieta Kubera, Agnieszka Kubik-Komar, Krystyna Piotrowska-Weryszko, Agata Konarska
University of Life Sciences in Lublin, Poland; elzbieta.kubera@up.lublin.pl

Pollen monitoring carried out using volumetric traps is a complex and time-consuming task. The final goal of our research is to create a system for automatically recognizing and counting pollen grains of individual taxa from microscopic images.

In this work, we compare three types of object detectors in terms of the recognition correctness of *Alnus*, *Betula*, *Corylus,* and *Carpinus* pollen grains - two one-stage detectors: YOLOv5 (in two versions) and RetinaNet, and a two-stage detector - Faster RCNN. Our dataset contains microscopic photos of the reference material. Individual images show only pollen grains of one taxon, thus we were able to avoid errors during the dataset annotation. Each detector model was built three times, so 12 models were obtained (4 detectors x 3 repetitions). The detector's training lasted for 500 epochs and consisted of fine-tuning some pre-trained models on our dataset.

We used the PyTorch library to build the Faster RCNN and RetinaNet models instead of the Detectron2 framework used in our previous studies. This change enables us to choose the best model in the training process based on its evaluation using the validation set. Therefore the final models were the best-rated models on the validation set.

When recognizing pollen grains for counting purposes, precision is crucial. In contrast, the standard metric of detection quality - mAP (mean Average Precision, which takes into account location errors) is less important in this case. Therefore, we used classification quality measures to evaluate each model, with particular emphasis on precision. Accordingly, we consider two types of YOLOv5 detectors, which differ in the model's fitness measure, used both at the training and evaluation stages. In addition to the standard fitness metric expressed by mAP, we propose considering only the precision and recall at 70% and 30%, respectively.

The same test set consisting of reference pollen pictures was used to compare the quality of each detector.

Values of classification measures were compared using a nonparametric rank-based alternative for ANOVA with repeated measures. We used the F1-LD-F2 design with taxon as a whole-plot factor(4 levels) as well as the detector (4 levels) and its repetition (3 levels) as the first and second sub-plot factor variables, respectively. In addition, multiple comparisons with the Bonferroni adjustment were applied.

The obtained results allowed us to indicate the YOLO detector as preferable to Faster RCNN and RetinaNet regarding all classification measures. There were no differences in the distributions of these measures between default and modified YOLO models. Additionally, there were no preferences in specific taxon classification precision by any of the studied detectors. However, we found that the recall distribution for investigated taxa differs significantly regarding final detectors. RetinaNet and Faster RCNN more often omitted partially visible pollen grains located near the picture border than YOLO detectors, which probably resulted in a difference in recall results.

Monday, 04/Sept/2023 2:20pm - 2:40pm
ID: 184 / **S11: 2**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Safety and benefit/risk assessment in master protocols
Keywords: Aggregate Safety Assessment Planning, Benefit-Risk Assessment Planning, Interdisciplinary collaboration

How to master the challenges of safety and benefit/risk assessment planning?

Jürgen Kübler
Quantitative Scientific Consulting, Germany; juergen.kuebler@qscicon.com

Aggregate safety assessment planning (ASAP) has recently be proposed (Hendrickson et al 2021). ASAP provides a guide for systematic product-level safety planning, standardized data collection and analyses, knowledge gap assessment and safety related communications. The concept can be extended to Benefit Risk Assessment Planning by taking efficacy and tradeoff between efficacy and safety into consideration. Both activities require an interdisciplinary collaboration along the life-cycle of a medical product.

One the of key features of both ASAP and BRAP is the understanding that this planning is not sufficient at the clinical trial level. Both a comprehensive safety assessment and the benefit-risk assessment require an ordered development programme that systematically addresses knowledge gaps. Master protocols by definition deal with generation and use of information across trials while not necessarily focusing on a single compound and/or single indication. Similar to the traditional approach to drug development, the discussion of the value of master protocols seems to focus on efficacy.

This presentations explores challenges and opportunities when extending the current concepts of ASAP and BRAP to master protocol.

References

1. Hendrickson, B.A., Wang, W., Ball, G. et al. Aggregate Safety Assessment Planning for the Drug Development Life-Cycle. Ther Innov Regul Sci 55, 717–732 (2021). https://doi.org/10.1007/s43441-021-00271-2
2. Bronson A, Chase MK, Fisher K, Millar D, Perlmutter J, Richardson N. Mobilizing the clinical trial ecosystem to drive adoption of master protocols. Clin Trials. 2022 Dec;19(6):690-696. doi: 10.1177/17407745221110199. Epub 2022 Sep 9. PMID: 36086812; PMCID: PMC9679560.
3. European Medicine Agency. Complex clinical trials – Questions and answers. 2022. https://health.ec.europa.eu/system/files/2022-06/medicinal_qa_complex_clinical-trials_en.pdf (assessed om 20 Feb 2023)
4. US Food and Drug Administration. 2022. Master Protocols: Efficient Clinical Trial Design Strategies to Expedite Development of Oncology Drugs and Biologics.
5. https://www.fda.gov/media/120721/download (assessed om 20 Feb 2023)

Monday, 04/Sept/2023 5:10pm - 5:30pm
ID: 403 / **S20: 4**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …)
Keywords: functional regression, functional quantile regression, empirical distribution, prediction sets

Estimating the conditional distribution in functional regression problems

Thomas Kuenzer[1], **Siegfried Hörmann**[2], **Gregory Rice**[3]
[1]Medical University of Graz, Austria; [2]Graz University of Technology, Austria; [3]University of Waterloo, Canada;
thomas.kuenzer@medunigraz.at

We consider the problem of consistently estimating the conditional distribution of a functional data object given functional covariates, assuming that the response and the predictor are related by a functional regression model. A nonparametric method is proposed that is based on the empirical distribution of the estimated model residuals. In the case of functional linear regression, consistent estimation of the conditional distribution can be achieved. This permits to describe interesting path properties of the response in a simple way. The usefulness of the method is demonstrated using both simulated and clinical data.

Tuesday, 05/Sept/2023 2:20pm - 2:40pm
ID: 290 / **S32: 2**
Presentation Submissions - Featured Session
Featured Sessions: Young Statisticians Sessions and Panel Discussion
Keywords: interpretable reporting, individualized treatment decisions, marginal effects, adjusted predictions

A flexible framework for interpretable and individualized reporting of model results

Hannah Kümpel, **Sabine Hoffmann**
LMU Munich, Germany; hannah.kuempel@ibe.med.uni-muenchen.de

The results of statistical models are frequently intended to inform evidence-based decision-making. However, commonly reported effect size measures like odds ratios lack interpretability and are often misunderstood. In the same way as statistical significance, they are, therefore, not very well suited when it comes to deriving practical decision rules. This particularly applies to situations where a medical practitioner tries to determine the best course of action by balancing costs, benefits, and uncertainties for patients based on individual characteristics. For example, the odds ratios from a logistic regression model can not be used to answer questions such as 'By how much does receiving treatment change the expected probability of heart failure for women between fifty and sixty years of age?'.

To facilitate evidence-based decision-making based on statistical analyses for medical practitioners, we propose a framework for individualizing model output post-inference. Specifically, we generalize the concepts of marginal effects and adjusted predictions to then define point estimates and uncertainty regions for both the average expected target variable given specific patient characteristics and the average expected absolute change resulting from changes in these characteristics. Along with these quantities, we propose corresponding visualization techniques that may be reported alongside classical effect size measures to improve the interpretability of study results.

Furthermore, we present a method to not only visualize average expected change but to take into account both estimation and sampling uncertainty by visually comparing the expected distribution of the target variable resulting from changes in patient characteristics.

A notable benefit of the proposed framework is that it allows for the specification of each defined quantity according to the exact research question at hand rather than having to adjust one's reporting as a function of the correct interpretation of a given effect size measure. This is achieved by using probability measures for averaging over predictor values and, furthermore, distinguishing between three axiomatic assumptions regarding the dependence structure of these predictors. We illustrate the proposed methodology by applying it to selected case studies in biomedical research, highlighting its practical relevance.

Monday, 04/Sept/2023 4:10pm - 4:50pm
ID: 489 / **S16: 1**
Presentation Submissions - Invited Session
Invited Sessions: Causal discovery with a view to the life sciences
Keywords: Causal discovery, Causal inference, Structure learning, Bayesian networks

Causal discovery: Benchmarking algorithms and Bayesian analyses in the life sciences

Jack Kuipers
ETH Zurich, Switzerland; jack.kuipers@bsse.ethz.ch

Probabilistic and causal graphical models are powerful tools to help understand and characterise complex mechanisms. Given a causal diagram, there are ways to estimate the effect of one variable on another from data. Causal discovery aims to find plausible causal relationships without prior knowledge of the graphical structure, a broad and fast-moving field. Selecting the most appropriate method from a plethora of available algorithms can be a daunting task. After surveying well-established approaches to structure learning, we present a workflow for large-scale benchmarking in a systematic and reproducible manner with the Benchpress platform. Its strengths include being fully modular and easily extendable to include new algorithms. In the life sciences, graphical models and structure learning may aid communication in complex scenarios and facilitate decision-making. For illustration, we discuss two case studies focusing on sampling-based methods, as they enable fully Bayesian analyses. These naturally characterise the uncertainty in the estimates of both network structures and their parameters. The first case study deals with uncovering patterns of genetic mutations in large-scale oncology datasets. As generative models, Bayesian networks offer a framework for model-based clustering. Integrating clinical covariates with the genomic profiles in the causal graphical model may provide more informative patient stratifications for developing personalised therapeutics. The second case study comes from psychiatric epidemiology, where we wish to estimate putative intervention effects from psychological survey data to better inform study design for interventional trials.

Monday, 04/Sept/2023 4:50pm - 5:10pm
ID: 208 / **S19: 3**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Clinical trial designs (adaptive and platform designs, external controls, …), Statistical modelling (regression modelling, prediction models, …)
Keywords: oncology, tumor growth, dose-response

Modeling the dose-response relationship using advanced tumor metrics

Cornelia Ursula Kunz, Stephan Lücke
Boehringer Ingelheim Pharma GmbH & Co. KG, Germany; cornelia_ursula.kunz@boehringer-ingelheim.com

The development of new drugs is often time-consuming and expensive – especially in oncology. Hence, there is a need to speed up the development process. Programs like the FDA's Critical Path Initiative in general and the Oncology Center of Excellence Project Optimus specifically focus on optimizing the dose selection process in oncology.

Unlike in other therapeutic areas, there is no formal dose finding in the sense of establishing a dose-response relationship and selecting an optimal dose. Typically, a dose (MTD /RP2D) is identified in Phase 1 and then carried forward to Phase 2 in which data on a binary clinical outcome like RECIST-based objective response is collected in a single-arm trial. If the response rate achieves some pre-defined criteria, the drug enters a Phase 3. One main question is whether other tumor measurements can be used in Phase 2 trials to better characterize the dose-response relationship. Tumor growth models which describe the change of the tumor burden over time in response to treatment using exponential models could provide alternative measures, as for example the g(rowth)-rate or the d(ecline)-parameter.

We investigate the mathematical properties of exponential tumor growth models and derive several equations and algorithms linking the g- and d-parameters to other tumor measures like response and progression as well as time-to-response, time-to-progression, and duration of response. The mathematical framework allows us to specify constrains like desired response rate, follow-up-time, and median time-to-response yielding unique solutions for the mean of the logarithm of the g- and d-parameter. Based on this, the framework can be used to jointly simulate response and time-to-event endpoints in oncology. In addition, it can be used to investigate the advantages and disadvantages of using the g- and d-parameter instead of the response rate for establishing a dose-response relationship in Phase 2.

Wednesday, 06/Sept/2023 9:10am - 9:30am
ID: 299 / **S47: 3**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Statistical software engineering in the pharmaceutical industry: Increasing productivity, transparency, and reproducibility via open source collaboration

To package or not to package - a pragmatic approach to deciding whether an R package is the right solution for your problem and alternatives to consider

Kevin Kunzmann
Boehringer-Ingelheim, Germany; kevin.kunzmann@boehringer-ingelheim.com

In recent years, statisticians and analysts have increasingly adopted open-source software such as the R language. The success of R in statistics and beyond is in large part due to its extensive ecosystem of open-source extension packages. Writing R packages is thus an increasingly important skill for statisticians to bring novel methodology into practice or when trying to make their workflows more reproducible and effectively share code with collaborators.

This talk addresses some potential downsides of wrapping R code in a package, like the burden of continued maintenance of a package, and highlights alternative formats of sharing functionality with collaborators. Many software engineering best-practices for R packages can be used outside of a full R package enabling a more adequate compromise between quality and simplicity depending on the application.

In this talk, the role of scripts, functions, literate programming (R markdown and Quarto), R packages, application programming interfaces (API), and shiny apps in the R ecosystem are reviewed and guidance on how to select he right tool for a particular objective is given. Hands-on recommendations on where and how software engineering best-practices like version control, testing, or documentation can be implemented outside of an R package context are discussed.

Knowledge of the broad range of options the R ecosystem offers for making statistical analysis code available can both lower the entry hurdle to newcomers to the R language and further increase the impact of more advanced R users.

Wednesday, 06/Sept/2023 11:00am - 11:20am
ID: 268 / **S52: 2**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Machine Learning and Data Science
Keywords: random forests, most representative trees, interpretable machine learning, clustering

Identifying different tree types based on clustering in random forests

Björn-Hergen Laabs, Lea Louisa Kronziel, Ana Westenberger, Inke R. König
Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck, Germany; b.laabs@uni-luebeck.de

Since the popularity of machine learning algorithms is ever increasing, methods for opening these black box methods become more and more important. In the case of random forests (RF), most representative trees (MRTs) have shown a big potential to facilitate the interpretation of the complex tree ensembles. The idea of MRTs is that the complete RF is represented by a single selected (S-MRT) or artificially generated tree (A-MRT). Due to their complex structure, it is likely that a single MRT is not able to capture the structure of all trees in the ensemble, especially when the trees in the RF are very diverse. This could, for instance, be the case with latent subgroups in the data, leading to trees that are specific for different subgroups.

Therefore, we propose a two-step procedure that firstly clusters the trees within a RF into different types of trees using a standard cluster algorithm and secondly generates a single MRT for each of the resulting clusters of trees. Thus, we end up with a small ensemble of clustered MRTs (C-MRT), that is better able to cover the diversity of the complete RF. Combined with the methods to obtain the single MRT this leads to either clustered selected MRTs (CS-MRT) or clustered artificial MRTs (CA-MRTs).

In an extensive simulation study, we will compare CA-MRTs and CS-MRTs with the previously described S-MRTs and A-MRTs regarding their prediction performance, ability to condense the information of the ensemble and coverage of the meaningful predictors. We simulate a standard setting including fixed main and interaction effects in high dimensional data where C-MRTs to proof that they are not inferior to normal MRTs as well as a setting where latent subgroups are present in the training data, which should lead to more diverse trees in the RF. Here C-MRTs should clearly outperform standard MRTs.

Additionally, we apply all methods to a genetic data set of X-linked dystonian-parkinsonism (XDP) and discuss the resulting MRTs with regard to recent results on genetic modifiers of age at onset in XDP.

Finally, we will add the new methods to our existing R package timbR (https://github.com/imbs-hl/timbR).

Thursday, 07/Sept/2023 9:30am - 9:50am
ID: 280 / S61: 3
Presentation Submissions - Invited Session
Invited Sessions: Statistical strategies in toxicology
Keywords: Genetic toxicology, validation

Statistics in a validation process in toxicology

Tina Lang
Bayer AG, Germany; tina.lang@bayer.com

During the early phases of drug development, the assessment of safety for candidate compounds is vital for any further progress. A guideline-driven well-established apparatus of tests is in place to check the different aspects of safety and toxicity (guidelines, e.g., for micronucleus test in vitro OECD Guideline 487 (2016) and ICH S2(R1) (2012)).

These guidelines focus primarily on the experimental conduct of the preclinical study and require the adherence to GLP (good Laboratory Practice). Besides the experimental part, the analysis environment and data analysis procedures must also be validated according to the quality assurance procedures

We have been involved with our colleagues from both genetic toxicology and quality assurance in the establishment of such a validation process. In addition to the statistical evaluation, the validation and qualification also encompassed the statistical software and the IT infrastructure.

In this presentation we want to share our experience from the process validation, including the fact that “validation” has different meanings in different contexts. Thus, as it is often the case, starting with defining technical terms for common ground between statisticians, toxicologists and quality experts is an important base for the successful integration of statistical analysis in regulatory toxicology studies.

We will present our journey, obstacles, and the common goal we achieved as a team in the end.

References:

1. ICH S2(R1) (2012), Harmonized Tripartite Guideline, “Guidance on genotoxicity testing and data interpretation for pharmaceuticals intended for human use”.
2. OECD/OCDE TG 487 (2016): “OECD Guideline for the testing of chemicals – In Vitro Mammalian Cell Micronucleus Test”.

Tuesday, 05/Sept/2023 2:40pm - 3:00pm
ID: 414 / S32: 3
Presentation Submissions - Featured Session
Featured Sessions: Young Statisticians Sessions and Panel Discussion
Keywords: meta-science, pre-registration template, large-scale real-data benchmark study, binary classification, events per variable

Confirmatory studies in methodological statistical research: concept and illustration

F. Julian D. Lange[1,2], Anne-Laure Boulesteix[1,2]
[1]Institute for Medical Information Processing, Biometry, and Epidemiology, LMU Munich, Germany; [2]Munich Center for Machine Learning (MCML), Munich, Germany; julian.lange@ibe.med.uni-muenchen.de

Hypothesis-generating, exploratory research and hypothesis-testing, confirmatory research are both essential to progress in science. However, failing to separate the two types of research can lead to non-replicable results when exploratory findings are misperceived or intentionally presented as confirmatory. To transparently conduct strictly confirmatory analyses, the practice of publicly registering research plans before the data analysis has become increasingly popular. This process is called pre-registration. For a number of applied research fields and study types, templates to aid researchers in specifying sufficiently detailed plans are available. In the context of methodological statistical research, however, the distinction between exploratory and confirmatory studies has received little attention in the scientific literature so far. Consequently, there is no guidance available regarding the pre-registration of methodological research in particular. To address this gap, this work proposes an approach for a strictly confirmatory real-data study in this field and provides a corresponding pre-registration template for comprehensively planning such a study. The suggested approach is illustrated with a large-scale benchmark experiment, and its results roughly confirm the findings of an existing simulation study by van der Ploeg et al. (2014). Specifically, the illustration indicates that untuned random forests (a) require more events per variable (EPV) than logistic regression to realize their predictive performance potential and (b) are prone to overfitting even when generated with a large number of EPV. It also demonstrates how pre-registration can prevent selective reporting and over-optimistic conclusions, thereby suggesting that the adoption of the proposed approach could lead to more credible methodological statistical research.

References

van der Ploeg, T., Austin, P. C., and Steyerberg, E. W. (2014). Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Medical Research Methodology*, 14:137.

Thursday, 07/Sept/2023 9:30am - 9:50am
ID: 285 / S63: 4
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), Machine Learning and Data Science, High dimensional data, genetic and x-omics data
Keywords: boosting, interactions, SNPs, decision trees, regression

Boosting interaction tree stumps for modeling gene–gene and gene–environment interactions

Michael Lau[1,2], Tamara Schikowski[2], Holger Schwender[1]
[1]Mathematical Institute, Heinrich Heine University, Düsseldorf, Germany; [2]IUF – Leibniz Research Institute for Environmental Medicine, Düsseldorf, Germany; michael.lau@hhu.de

The development of complex phenotypes often not only depends on isolated genetic and environmental risk factors but also on their interplay. These phenomena are known as gene–gene (GxG) and gene–environment (GxE) interactions. A GxG interaction is present if the effect of participating loci depends on the presence of other participating loci, while a GxE interaction is defined as different susceptibilities to an environmental risk factor depending on the genotype. Classical procedures for modeling phenotypes based on genetic risk factors such as SNPs (single nucleotide polymorphisms) either depend on simplifying assumptions such as linearity in generalized linear models or produce non-interpretable black-box models such as random forests or deep neural networks.

To overcome these drawbacks, we propose a statistical learning method called BITS (boosting interaction tree stumps) that aims at fitting simple-to-read linear models that incorporate GxG and GxE interactions. In every boosting iteration, tree stumps are fitted that – instead of the usual split on a single input variable – may split on interactions of the input variables. To avoid unnecessarily complex models, these interaction tree stumps are regularized for including long interactions and the resulting model is pruned and transformed into a linear model using the elastic net. GxE interactions are incorporated by including the environmental variable and potential interactions with the identified terms.

In contrast to many related methods, the computational complexity of BITS scales linearly with the number of input variables such that BITS is also suited for high-dimensional tasks.

In a simulation study, it is shown that BITS outperforms existing methods regarding the predictive ability on unseen data. Moreover, multisplitting is employed for statistically testing GxG and GxE interactions. The simulations also show that BITS controls the type I error rate for detecting GxG and GxE interactions, true underlying terms are often identified, and GxE interactions are detected with a high power. Furthermore, BITS and related methods are applied and compared in a real data application analyzing data from a German cohort study.

ID: 484 / **STRATOS 1: 3**
Presentation Submissions - Featured Session
Featured Sessions: Ten years of the STRengthening Analytical Thinking for Observational Studies (STRATOS) initiative – progress and looking to the future
Keywords: Missing data; Sensitivity analysis; Guidance

Level 1 guidance on conducting and reporting sensitivity analyses for missing data

Katherine Lee[1], Rheanna Mainzer[1], James Robert Carpenter[2,3,4]
[1]Murdoch Children's Research Institute, Melbourne, Australia; [2]London School of Hygiene & Tropical Medicine, United Kingdom; [3]MRC CTU at UCL, London UK; [4]for TG1; james.carpenter@lshtm.ac.uk

Missing data are common in observational studies. When estimating a target parameter in the presence of missing data, the researcher (either implicitly or explicitly) makes assumptions about the unknown missingness mechanism.

An important, but often overlooked step of the analysis is examining the robustness of estimates and inferences to alternative plausible assumptions about the missingness mechanism, and in particular conducting analyses that allow the missingness to depend on the missing values themselves, sometimes referred to as a "missing not at random" or a "delta-adjusted" analysis.

We previously outlined a framework for handling and reporting the analysis of incomplete data in observational studies where we encourage researchers to think systematically about missing data and transparently report the potential effect on the study results[1]. In this talk, we extend this framework to the planning, conduct and reporting of sensitivity analyses which incorporate external information about how the missing values differ to those observed.

We illustrate the process using a case study from the Avon Longitudinal Study of Parents and Children, providing practical guidance that can be tailored to the problem at hand. We hope this guidance will make such sensitivity analyses more accessible to researchers, increasing its use in practice, and increasing the confidence in research findings from incomplete data.

Reference:

Lee, K. J., Tilling, K. M., Cornish, R. P., Little, R. J. A., Bell, M. L., Goetghebeur, E., Hogan, J. W. and Carpenter J. R. (2021) Framework for the treatment and reporting of missing data in observational studies: The Treatment And Reporting of Missing data in Observational Studies framework. Journal of Clinical Epidemiology, **134**, 79-88.

Thursday, 07/Sept/2023 9:10am - 9:30am
ID: 243 / **S63: 3**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), Machine Learning and Data Science, High dimensional data, genetic and x-omics data, Real world data and evidence
Keywords: multi-omics data, prediction, benchmark study, survival analysis, cancer

Multi-omics data integration: Does more mean better for predictive modeling? A large-scale benchmark study

Yingxia Li[1], Ulrich Mansmann[1], Roman Hornung[1,2]
[1]Institute for Medical Information Processing, Biometry and Epidemiology, University of Munich; [2]Munich Center for Machine Learning (MCML); hornung@ibe.med.uni-muenchen.de

Predictive modeling based on multi-omics data, that is, several types of omics data available for the same patients, has demonstrated the ability to potentially outperform single-omics predictive modeling. Previous research on using multi-omics data for prediction has focused on combining many types of data. However, collecting many omics data types is complex and costly, which is why it would be beneficial to collect only those omics data types that contribute to improving predictive performance. It is, however, unclear which combinations of omics data types are most effective and which types can generally be omitted without compromising predictive performance.

We compared the predictive performance of all 31 possible combinations of five genomic data types using different prediction methods applied to 14 cancer datasets with survival outcome. The data types considered were mRNA, miRNA, methylation, mutation, and copy number variation data. Clinical data were included and prioritized in each prediction model. To investigate the stability of the results, bootstrap analysis was performed at the level of the included datasets.

Contrary to our expectations, combining larger numbers of omics data types tended to degrade predictive performance. Instead, using only mRNA data or a combination of mRNA and miRNA data was sufficient in most cases. Although the number of datasets included in our study is comparatively large, it is still limited, which is why our results must be interpreted with caution. Nevertheless, they strongly suggest that integrating many omics data types in a multi-omics prediction context may be counterproductive.

Monday, 04/Sept/2023 3:00pm - 3:20pm
ID: 231 / **S13: 4**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …)
Keywords: Modelling, Nonlinear Model, Confidence Intervals, Profile Likelihood, Validation

Uncertainty Estimation in Nonlinear Models within the Profile Likelihood Framework

Tim Litwin, Clemens Kreutz
Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center – University of Freiburg, Germany;
tim.litwin@uniklinik-freiburg.de

The use of nonlinear models has become increasingly popular in many quantitative sciences. Specifically, such non-linear models are applied in the life sciences, because non-linear behavior in living systems arises from the existence of complex networks of interactions and feedback loops.

From a mathematical point of view, nonlinear models can exhibit a variety of features, which do not occur in linear models. However, applying statistical methods developed in the linear context on non-linear models is still common-practice. This is usually motivated by local approximation of the nonlinear model, referring for example to asymptotic properties of Wald confidence intervals based on large sample sizes, which is often insufficient in the finite sample size case.

In this talk, we recommend the profile likelihood method as a means to construct confidence intervals as a viable alternative to classical linear approaches. The profile likelihood approach branches into multiple different methods suitable for different contexts, which can be broadly categorized into uncertainty estimation and experimental design. The profile likelihood can accurately capture the non-linear model behavior in the estimation of confidence intervals for parameters and predictions. Specifically, the method lends itself to detect parameters with associated non-finite confidence intervals which arise from model overparametrization or lacking data quality. Incidentally, evaluation of the profile likelihood can additionally be used to propose and verify informative experimental designs for these poorly identified parameters. The basic profile likelihood approach is implemented in multiple modelling frameworks, making numerical evaluation feasible and easy to use and interpret.

We conclude that the profile likelihood approach should be considered as a standard tool in the analysis of uncertainties in parameters and predictions of non-linear models. Therefore, this talk sets out the profile likelihood approach as a both theoretically well-founded as well as a practically feasible alternative to classical linear approaches, aiming to familiarize the audience with the concept.

Monday, 04/Sept/2023 2:40pm - 3:00pm
ID: 464 / **S12: 3**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Critical cross-disciplinary collaboration in dose optimization in oncology
Keywords: Project Optimus, Dose selection

Pragmatic and Holistic Approach for Dose Finding and Optimization in Oncology Drug Development – A Clinical Pharmacology Point of View on Project Optimus

Jiang Liu
US FDA, United States of America; jiang.liu@fda.hhs.gov

Efficient dose finding and optimization in general population and an individual patient have always been a critical issue of drug (including oncology drug) development. With advances in cancer biology and new molecular targeted agents and immunotherapies, pressing need for an improved oncology drug dose finding and optimization strategy has been highlighted, as demonstrated particularly by the recent Project Optimus from the FDA. The new strategy will require a multi-disciplinary collaboration to integrate all available nonclinical and clinical information, including pharmacokinetic (PK), pharmacodynamic (PD), activity or efficacy, safety and tolerability, and patient-disease-trial factor impacts and an understanding of dose- and exposure-response relationships for safety and efficacy. This presentation will provide an overview of current landscape of dose finding and optimization in the oncology drug development. A pragmatic and holistic approach integrating the totality of evidence at different stage of drug development for dose selection and optimization for general or specifical population will be discussed and illustrated using representative case examples.

Monday, 04/Sept/2023 5:10pm - 5:30pm
ID: 164 / S17: 4
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Prognostic and predictive biomarkers in personalized medicine

Logic respecting efficacy measures in the presence of prognostic or predictive biomarker subgroups

Yi Liu[1,3], **Hong Sun**[2,3]
[1]Nektar Therapeutics; [2]Bristol Myers Squibb; [3]Oncology Estimand Working Group Task Force 8; hong.sun@bms.com

In the era of precision medicine, understanding treatment effect in biomarker defined subgroups in relationship with overall population is essential. For continuous outcomes, Least Square estimates include an interaction term to enable an unbiased estimation of treatment effect in the overall population (overall treatment effect) by linearly combining treatment effects of the two complementary subgroups. Such logic is carried to binary and time-to-event outcomes models in most statistical software where model parameters are linearly combined in the log scale and then exponentiated to represent overall treatment effect. Although guaranteeing logical inference in appearance, such calculations do not correspond to the true overall treatment effect which may in fact be illogical for efficacy measures such as odds ratio and hazard ratio, i.e., the overall treatment effect is outside the range of subgroups effects. To correctly derive efficacy in the overall population, a principle called Subgroup Mixable Estimation (SME) should be followed. We illustrate these common mistakes and demonstrate the application of SME using real trial data.

Thursday, 07/Sept/2023 9:10am - 9:30am
ID: 364 / S58: 4
Presentation Submissions - Regular Session (Default)
Topic of Submission: Clinical trial designs (adaptive and platform designs, external controls, …)
Keywords: platform trial, response-adaptive randomisation, melanoma, immunotherapy

A biomarker-guided Bayesian response-adaptive phase II trial for patients with metastatic melanoma: The Personalized Immunotherapy Platform (PIP)-Trial design

Serigne N. Lo[1,2,3], Tuba N. Gide[1,2,3], Nurudeen Adegoke[1,2,3], Yizhe Mao[1,2,3], Monica Lennox[1,2,3], Saurab Raj Joshi[1,2,3], Camelia Quek[1,2,3], Ismael A. Vergara[1,2,3], Nigel Maher[1,2,3,4], Alison Potter[1,2,3,4], Robyn P.M. Saw[1,2,6,7], John F. Thompson[1,2,6,7], Andrew J. Spillane[1,2,5,7], Kerwin F. Shannon[1,2,7,8], Matteo S. Carlino[1,2,9], Maria Gonzalez[1], Alexander M. Menzies[1,2,5,7], Inês Pires da Silva[1,2,3,9], Stephane Heritier[10], Richard A. Scolyer[1,2,3,4], Georgina V. Long[1,2,3,5,7], James S. Wilmott[1,2,3]

[1]Melanoma Institute Australia, The University of Sydney, Sydney, NSW Australia; [2]Faculty of Medicine and Health, The University of Sydney, Sydney, NSW Australia; [3]Charles Perkins Centre, The University of Sydney, Sydney, NSW Australia; [4]Royal Prince Alfred Hospital and NSW Health Pathology, Sydney, NSW, Australia; [5]Royal North Shore Hospital, Sydney, Australia,; [6]Royal Prince Alfred Hospital, Sydney, NSW Australia; [7]Mater Hospital, North Sydney, NSW Australia; [8]Chris O'Brien Lifehouse, Sydney, NSW Australia; [9]Westmead and Blacktown Hospitals, Sydney, NSW Australia; [10]Monash University, Melbourne, VIC Australia; Serigne.Lo@sydney.edu.au

Anti-PD1-based immunotherapies have been approved for many cancer types and are now front-line treatment for patients with metastatic melanoma. Despite this, about 50% of these patients fail to respond to therapy. It is therefore critical to identify and prioritise patients with a low likelihood of response to front-line treatment and be able to investigate alternative effective therapies within clinical trials. With this background, we designed the phase II Personalised Immunotherapy Platform Trial (PIP-Trial), an investigator-initiated clinical trial evaluating two biomarker-driven treatment selections of 5 novel agents as: 1) first-line therapy in metastatic melanoma for patients predicted to be resistant to Australian Government subsidised standard drug therapies (Part A), and 2) second-line therapy in patients who experienced disease progression after receiving standard PBS therapies (Part B). Part A is a Bayesian adaptive multi-arm multi-stage design using response-adaptive randomisation after a burn-in period where patients are randomised to the existing arms with equal probability. Thereafter, interim analyses will be carried out to determine continuation or discontinuation of each arm based on their performance. Part B is designed using OPTIM-ARTS (Open Platform Trial Investigating Multiple Compounds – Adaptive Randomized Design with Treatment Selection). The design consists of a selection and an expansion phase to identify subgroups of patients for whom a novel agent works best as second-line therapy (without control). The primary outcome is 6-month RECIST objective response (ORR). Individual treatment arms will be halted when the posterior probability of observing a clinically-significant effect on the primary outcome (i.e. 6-month ORR) is below a pre-defined threshold.

The operational characteristics of the design were investigated through simulations considering various plausible scenarios. These show good performance of the design and a better allocation of resources for a reasonable maximum patient sample size of 216. All simulations were based on the R package BATS.

Tuesday, 05/Sept/2023 12:20pm - 12:40pm
ID: 418 / **S23: 5**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Time-to-Event Analysis
Keywords: Stratified Cox mode, partly interval censoring, maximum penalized likelihood

Penalized likelihood estimation of stratified semi-parametric Cox models under partly interval censoring

Jun Ma
Macquarie University, Australia; jun.ma@mq.edu.au

In survival data analysis, the stratified Cox model becomes a popular option when the proportional hazards assumption of the conventional Cox model does not hold for certain covariates. When survival times include interval-censored observations, the method of maximum partial likelihood is not viable, and thus cannot be applied. We present a penalized likelihood method for estimating the model parameters, including the baseline hazards. Penalty functions are used to produce smoothed baseline hazards estimates, and also to relax the requirement on optimal number and location of the knots used in the baseline hazards estimates. We also explain a large sample normality result for the estimates, which can be used to make inferences on quantities of interest, such as survival probabilities, without relying on computing-intensive resampling methods.

Tuesday, 05/Sept/2023 2:40pm - 3:00pm
ID: 244 / S34: 3
Presentation Submissions - Regular Session (Default)
Topic of Submission: Meta-Analysis and Systematic Reviews
Keywords: meta-analysis, missing data, multiple imputation, heterogeneity, finite mixture model

Imputation of informatively missing data in meta-analyses

Christine Macare[1], Thomas Lehmann[2], Peter Schlattmann[2]
[1]Heidelberg University, Germany; [2]University Clinic Jena, Jena University, Germany; christine.macare@hotmail.com

Background: Decision making for national health care relies on summaries of individual clinical studies. Meta-analyses offer a natural way to combine findings and quantify pooled treatment effects estimates. However, input data to meta-analyses is frequently hampered by missing values and heterogeneity between individual studies. Multiple imputation (MI) is a popular approach to handling missing data in medical research as evidenced by current Cochrane guidelines, yet little is known about its applicability for estimating pooled treatment effects with simultaneous estimation of heterogeneity variance. Proper specification of the imputation model is needed in order to avoid biased parameter estimates.

Methods: Meta-analytical data and missingness according to a previously specified non-response mechanism covering different non-response rates were simulated (see e.g. Sidik & Jonkman, 2005; Lehmann & Schlattmann, 2017). Performance of random effects modelling of complete cases, mixture modelling of complete cases, random effects modelling in combination with imputed case analysis (ICA, Higgins, White & Wood, 2008) and mixture modelling with simultaneous MI for the estimation of pooled treatment effects and heterogeneity variance were contrasted. Estimation performance was evaluated using bias and mean square errors (MSE) and respective Monte Carlo SE.

Results: Main results indicated the benefit of using an imputation-based method (MI in combination with mixture modelling and in some cases ICA) for the performance of the estimation of bias and MSE for the pooled treatment effects and the heterogeneity variance. Performance was affected by the non-response rate, even at moderate rates (e.g. > 0.3). Complete case analysis in combination mixture modelling showed acceptable performance with regard to bias and MSE at low rates of non-response.

Conclusion: Overall, the current findings highlight the diversity in analysis strategies for handling missing data in meta-analyses and support the use of imputation-based strategies for achieving a performance of effect estimation that is characterised by acceptable bias and MSE.

Wednesday, 06/Sept/2023 11:20am - 11:40am
ID: 210 / **S50: 2**
Presentation Submissions - Invited Session
Invited Sessions: Covariate Adjustment in RCTs: Translating theory into practice within a pharmaceutical company via a data challenge
Keywords: Covariate adjustment, data challenge

Organizing a Data Challenge on Covariate Adjustment in RCTs

Dominic Magirr
Novartis, Switzerland; dominic.magirr@novartis.com

Covariate adjustment in randomized trials is a currently a topic of interest for health authorities, highlighted by a recent FDA guidance document and EMA qualification opinion on a particular type of adjustment.

How should pharmaceutical companies react to these developments? Are new analysis techniques required? How can awareness and comprehension of new methods be spread across a large organization? To help address these questions, an internal “Covariate Adjustment Challenge” was run within the Analytics department at Novartis. 23 participating teams were given access to data from five prior studies in a particular indication. The aim was to propose either a single "super" covariate or a pre-specified set of baseline covariates that could be used in covariate-adjusted analyses of key endpoints in a subsequent study in the same indication. Upon the “test data” becoming available, teams were scored according to the gain in precision from their proposed adjustment compared to an unadjusted analysis.

This talk will cover the background to the project, the scoring metrics, as well as a high-level overview of the results.

Monday, 04/Sept/2023 5:10pm - 5:30pm
ID: 294 / **S16: 3**
Presentation Submissions - Invited Session
Invited Sessions: Causal discovery with a view to the life sciences

Causality-inspired ML: what can causality do for ML?

Sara Magliacane
University of Amsterdam, Netherlands, The; sara.magliacane@gmail.com

Applying machine learning to real-world cases often requires methods that are robust w.r.t. heterogeneity, missing not at random or corrupt data, selection bias, non i.i.d. data etc. and that can generalize across different domains. Moreover, many tasks are inherently trying to answer causal questions and gather actionable insights, a task for which correlations are usually not enough. Several of these issues are addressed in the rich causal inference literature. On the other hand, often classical causal inference methods require either a complete knowledge of a causal graph or enough experimental data (interventions) to estimate it accurately. Recently, a new line of research has focused on causality-inspired machine learning, i.e. on the application ideas from causal inference to machine learning methods without necessarily knowing or even trying to estimate the complete causal graph.

In this talk, I will present an example of this line of research in the unsupervised domain adaptation case, in which we have labelled data in a set of source domains and unlabelled data in a target domain ("zero-shot"), for which we want to predict the labels. In particular, given certain assumptions, our approach is able to select a set of provably "stable" features (a separating set), for which the generalization error can be bound, even in case of arbitrarily large distribution shifts. As opposed to other works, it also exploits the information in the unlabelled target data, allowing for some unseen shifts w.r.t. to the source domains. While using ideas from causal inference, our method never aims at reconstructing the causal graph or even the Markov equivalence class, showing that causal inference ideas can help machine learning even in this more relaxed setting.

Monday, 04/Sept/2023 5:10pm - 5:30pm
ID: 297 / **S21: 4**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), Prevention and handling of missing data, Epidemiology
Keywords: population attributable fraction, average sequential population attributable fraction, Monte Carlo simulation, sample size estimation

Performance of point and interval estimators for average sequential attributable fraction - A simulation study

Carolin Malsch
University of Greifswald, Germany; carolin.malsch@uni-greifswald.de

Population-attributable fraction (PAF) is used in epidemiology, health services and public health research to prioritize targets for intervention programs. It allows to quantify, from the population perspective, the overall effect of a set of risk factors on an outcome of interest. It takes the risk factor's effect size as well as its prevalence into account. The average sequential PAF (asPAF), a partialization approach carrying desirable properties, is increasingly used in practical applications recently. The presented simulation study characterizes non-model-based and model-based estimators of the asPAF and Wald and Monte Carlo type confidence intervals.

A good performance of point estimators depends on (a) completeness of variables in the model, and (b) the correct specification of the regression formula with respect to interaction terms in case of model-based estimation. The model-based estimator outperforms the non-model-based estimator and is unbiased even in small samples and in situations with small outcome prevalence. However, computational time of the model-based estimator increases rapidly with increasing sample size and number of variables considered. Resampling-based confidence estimators such as Bootstrap with normality assumption and percentile as well as Jackknife are suited for confidence interval estimation. Here, the computational time especially in conjunction with the model-based estimator increases super-linear with increasing sample size.

Sufficient sample sizes assuring a desirable performance of asPAF exceed those for relative effect measures such as relative risk and odds ratio noticeably. Further, the required sample size increases with a higher number of variables and a lower prevalence of outcome. While sample sizes for PAF are known to decrease with increasing prevalence and effect size of a risk factor, a reverse relation can be observed for asPAF. Sample size estimation can be conducted using Monte Carlo simulation for every conceivable scenario.

Wednesday, 06/Sept/2023 11:00am - 11:20am
ID: 272 / **S53: 2**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Real world data and evidence
Keywords: metascience, researcher degrees of freedom, replicability, stability

Quantifying and comparing the impact of different sources of uncertainty in the analysis of electronic health records: An application to intraoperative hyperoxemia

Maximilian Michael Mandl[1,5], Andrea Becker-Pennrich[1,3], Ludwig Christian Hinske[3,4], Sabine Hoffmann[1,2], Anne-Laure Boulesteix[1,5]

[1]Institute for Medical Information Processing, Biometry, and Epidemiology, Ludwig-Maximilians-Universität München; [2]Department of Statistics, Ludwig-Maximilians-Universität München; [3]Department of Anesthesiology, Ludwig-Maximilians-Universität München; [4]Institute for Digital Medicine, University Hospital of Augsburg; [5]Munich Center for Machine Learning (MCML); mmandl@ibe.med.uni-muenchen.de

In recent years, the scientific community has become aware of the fact that there is high analytical variability when different researchers study the same research question on the same data set. If this phenomenon is combined with selective reporting, it may lead to an increased rate of false positive results, inflation in effect sizes, and overoptimistic measures of predictive performance. Hoffmann et al. (2021) argued that this analytical variability can be explained by six sources of uncertainty that are omnipresent in empirical research regardless of the respective discipline: sampling, measurement, model, parameter, data pre-processing, and method uncertainty. Failure to take this variety of uncertainties into account may lead to unstable, supposedly precise, and overoptimistic results which ultimately results in research findings that are not replicable. In the long run this has devastating consequences on the credibility of research findings, in particular if independent teams of researchers publish contradicting results on the same data set.

In the field of medicine, the increasing accessibility of large data sets that were not originally collected for research objectives – such as electronic health records or administrative claims data – elicits optimism and expectations of “real-world” evidence and individualized treatment protocols. Nevertheless, these optimistic prospects are paralleled by a mounting awareness that we are confronted with an even greater degree of choices in analyzing this type of data as opposed to conventional observational studies.

A recent study by Becker-Pennrich et al. (2022) used routinely collected data regarding adult craniotomies in order to develop a random forest regressor to impute perioperative paO2 values. These results may be applied to statistical inference for analyzing the impact of perioperative paO2 on, e.g., in-hospital mortality. Failure to consider the uncertainty of these estimates in the first-stage (i.e. imputation of paO2 values) may lead to multiple noncongruent results in the second-stage (i.e. statistical inference) of the analysis.

In the present project we extend this study in order to identify and quantify the different sources of uncertainty and thus assess the stability of results on a routinely collected data set – focusing on both machine learning methods and classical statistical inference in this nested analysis setting. Our focus lies on evaluating the impact of multiple choices made throughout the analysis (“researcher degrees of freedom”) including data pre-processing and model selection decisions.

References

1. Andrea S Becker-Pennrich, Maximilian M Mandl, Clemens Rieder, Dominik J Hoechter, Konstantin Dietz, Benjamin P Geisler, Anne-Laure Boulesteix, Roland Tomasi, and Ludwig C Hinske. Comparing supervised machine learning algorithms for the prediction of partial arterial pressure of oxygen during craniotomy. medRxiv, 2022. doi:10.1101/2022.06.07.22275483.
2. Sabine Hoffmann, Felix Schönbrodt, Ralf Elsas, Rory Wilson, Ulrich Strasser, and Anne-Laure Boulesteix. The multiplicity of analysis strategies jeopardizes replicability: lessons learned across disciplines. Royal Society Open Science, 8(4):201925, 2021.

Monday, 04/Sept/2023 2:00pm - 2:20pm
ID: 118 / **S12: 1**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Critical cross-disciplinary collaboration in dose optimization in oncology
Keywords: oncology drug development, dose-optimization, MTD

Designing Dose-Optimization Studies in Cancer Drug Development: Discussions with Regulators

Olga Marchenko[1], Rajeshwari Sridhara[2], Qi Jiang[3], Elizabeth Barksdale[4]
[1]Bayer, United States of America; [2]Oncology Center of Excellence, FDA, United States of America; [3]Seagen, United States of America; [4]LUNGevity Foundation, United States of America; olga.marchenko@bayer.com

The ASA Biopharmaceutical Section Statistical Methods in Oncology Scientific Working Group and the LUNGevity Foundation organized in coordination with the US FDA Oncology Center of Excellence four forums on dose-optimization studies. Discussions at these forums were focused on statistical considerations in designing dose-optimization studies of products for treatment of cancer patients at pre- and post-approval stages. Several appealing ideas and methods were proposed and discussed at these forums. Although a consensus was not achieved on every point, speakers and panelists agreed that we should use better design options and strategies. The main ideas and discussion points from these forums will be summarized and reviewed in this presentation.

Tuesday, 05/Sept/2023 3:00pm - 3:20pm
ID: 339 / S35: 4
Presentation Submissions - Regular Session (Default)
Topic of Submission: Epidemiology, Real world data and evidence, Statistical hypothesis testing (covariate adjustment, nonparametric methods, multiple comparisons, …)
Keywords: longitudinal studies, electronic health records, adverse health outcomes, atopic eczema, multiple comparisons

Adverse health outcomes among people with atopic eczema: a consistent application of longitudinal study design to multiple outcomes

Julian Matthewman[1], Sinéad Langan[1], Spiros Denaxas[2]
[1]London School of Hygiene & Tropical Medicine, United Kingdom; [2]University College London, United Kingdom;
julian.matthewman1@lshtm.ac.uk

Population-based cohort studies using longitudinal electronic health records (EHR) data are commonly used to explore adverse health outcomes for specific conditions (e.g., is eczema associated with subsequent development of fractures/cancer/cardiovascular disease/etc...). These studies aim to answer causal questions to provide actionable evidence for decision makers, however the status quo of conducting individual “one exposure – one outcome” studies is problematic, in its inefficiency, lack of transparency, lack of comparability between studies, and due to concerns about publication biases. Applying generic and adaptable approaches consistently to multiple research questions, while preserving the ability to incorporate the required expert knowledge and critical thinking, will allow generating evidence on important health related questions more quickly, make results more comparable, reporting more transparent, and studies more easily updateable with components such as more detailed disease phenotypes.

We describe how "hypothesis-testing" EHR studies are currently conducted and delivered as a "one exposure - one outcome" package, and problems with this approach. We then explore how a "one exposure - many outcomes" approach could address problems with current approaches and propose a framework to implement such an approach. We will discuss 3 main themes: 1. The role of heterogeneity between studies, i.e. which parts of studies can remain the same for research questions on different outcomes, and which parts should be different; 2. The role of clinical expertise and for each outcome selecting which analyses should be considered as main analyses and which as sensitivity analyses; 3. The organisation of outcomes into categories, based on universally applicable hierarchies, and exposure-specific categories.

We will demonstrate the approach using the applied example of adverse health outcomes in adults with (atopic) eczema. Eczema may be related to several adverse health outcomes, however for most previously explored outcomes there is only low or moderate certainty evidence, and some outcomes may be unknown. This uncertainty should be addressed in order to improve patient management including implementing screening and preventive measures where appropriate. We will describe a framework for a research pipeline, with generic methods to set up cohorts, estimate exposure-outcome associations within these cohorts, and account for confounding. Using this pipeline, we will explore adverse health outcomes associated with eczema, replicating previously published cohort studies, and investigating associations with outcomes that have not or not adequately been researched in the existing literature. The choice of outcomes will be guided by findings from a recent large-scale review of the previous evidence. Results for the applied example will be produced shortly, pending data access.

In summary, we will create a high-quality and comprehensive evidence source on the topic of adverse health outcomes associated with eczema while describing a framework adaptable to other skin diseases and other areas of research.

Wednesday, 06/Sept/2023 11:40am - 12:00pm
ID: 230 / **S55: 4**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Real world data and evidence, Software Engineering, Preclinical drug development, safety and toxicology
Keywords: Historical control data, bio-assay, bootstrap calibration, assay validation and qualification

The application of prediction intervals in pre-clinical statistics and toxicology using the R package predint

Max Menssen
Leibniz Universität Hannover, Germany; menssen@cell.uni-hannover.de

In several pre-clinical or toxicological applications, it is of interest to verify, if an actual observation is in line with historical knowledge. A strategy for this verification is the application of prediction intervals, that should cover an actual (or future) observation with a predefined probability.

For example, dichotomous endpoints such as tumor incidences obtained in longterm carcinogenicity assays lead to overdispersed binomial data. Counted observations such as the number of reverant bacteria colonies in the Ames assay can be modelled to be overdispersed Poisson. Continuous data (e.g. body weights or immunogenicity reactions) often involves one or several nested random effects, when several historical studies are jointly assessed.

Hence, prediction intervals for overdispersed binomial and Poisson data are implemented in the R package predint. Furthermore, prediction intervals computed based on random effects models, that allow for different random effect structures in the historical as well as in the actual data, are available via predint.

The use of the proposed prediction intervals will be demonstrated based on real life data.

Monday, 04/Sept/2023 4:30pm - 4:50pm
ID: 223 / **S19: 2**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Clinical trial designs (adaptive and platform designs, external controls, …), Statistical modelling (regression modelling, prediction models, …), Time-to-Event Analysis
Keywords: joint models, surrogacy, causality, cancer

Deconstructing PFS to understand the treatment effect and predict OS in oncology drug development

Francois Mercier, Georgios Kazantzidis, Daniel Sabanes-Bove, Beyer Ulrich
F. Hoffmann-La Roche, Switzerland; francois.mercier@roche.com

Oncology clinical development benefits from well-defined and widely accepted standards to evaluate anti-tumor drug activity. Indeed, the RECIST1.1 guideline offers a framework to harmonize the interpretation of changes in tumor burden for patients with solid tumors. To this aim, it brings various information together, namely radio imaging data on target and non-target lesions, as well as data on the emergence of new lesions, or clinical judgment. For target lesions, the sum of diameters (SLD) is calculated in every patient at regular time intervals during the trial. For active drugs and responder patients, the expectation is for SLD to decrease over time (tumor shrinkage); after a while, SLD is typically growing due to the loss of treatment effect. For inactive drugs or non-responder patients, SLD is monotonically increasing over time (tumor growth). The variability in individual longitudinal profiles can be studied using tumor growth inhibition (TGI) non-linear mixed-effect models.

Progression-free survival (PFS) is commonly used as a primary or secondary endpoint in oncology clinical trials. This right-censored variable is defined as the time to the earliest of either death (from any cause, OS) or disease progression (PD, per RECIST1.1). As such PFS is mixing events of different natures, occurring in sequence with PD preceding death. Instead of mingling them, we consider a joint model including (i) a TGI sub-model, (ii) a time-to-event OS sub-model, and (iii) a term measuring the strength and nature of the TGI-OS association. With this model, it becomes possible to investigate the effect of an intervention on each component of the bivariate process and to quantify the proportion of treatment effect on overall survival mediated by changes in SLD. We present an example showing the proportion of treatment effect mediated by SLD in urothelial cancer patients treated with atezolizumab. The model can also be used to predict OS based on early measures of SLD. Based on simulations, we illustrate how SLD can be used as a proxy for OS to support early decision-making. Extensions of this work are then discussed, including the potential role of non-target and new lesions in predicting OS.

Thursday, 07/Sept/2023 8:50am - 9:10am
ID: 146 / S59: 2
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Volume-outcome relationships in health care

The Assessment of Volume-Outcome Associations at IQWiG

Claudia-Martina Messow, **Jona Lilienthal**
Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG), Germany; claudia-martina.messow@iqwig.de

For a range of medical services provided, an association between the number of procedures carried out at a medical centre, i. e. the volume, and the outcomes of patients treated at that centre has been shown. Since 2003, the Federal Joint Committee (G-BA) sets binding minimum volume standards for certain planned inpatient services for a hospital to be permitted to deliver this service. When considering introducing a minimum volume standard, the G-BA can commission IQWiG with determining whether there is sufficient scientific evidence of any volume-outcome association with respect to this service. This presentation will give an overview of the methods used by IQWiG to assess the validity of studies investigating volume-outcome associations in healthcare.

As a first step, any relevant clinical evidence is identified in a literature search. Evidence from observational studies as well as controlled intervention studies (the intervention being the setting of a minimum volume) is considered. The studies have to meet a set of inclusion and exclusion criteria to ensure a minimum level of quality. In a next step, the selected studies are assessed for a range of quality aspects in order to rate their explanatory power. These include the quality of the data, details of statistical modelling, e.g. accounting for clustering, and the quality of reporting. Each reported result for any outcome relevant to the patient is subsequently assessed separately for its usability based on the analysis carried out. Then, the results for all relevant outcomes are extracted and compared. Due to the diversity of the studies included, results can generally only be synthesised qualitatively.

Monday, 04/Sept/2023 2:40pm - 3:00pm
ID: 216 / **S9: 2**
Presentation Submissions - Invited Session
Invited Sessions: A causal inference perspective on estimands in clinical trials
Keywords: Causal inference, extrapolation, estimand, randomized trials, positivity violation

The danger of extrapolation in RCTs and how to avoid it

Hege Michiels[1], **An Vandebosch**[2], **Stijn Vansteelandt**[1]
[1]Ghent University, Belgium; [2]Janssen R&D, Belgium; hege.michiels@ugent.be

When choosing estimands and estimators in randomized clinical trials, caution is warranted, as intercurrent events, such as, due to patients who switch treatment after disease progression, are often extreme. Statistical analyses may then easily lure one into making large implicit extrapolations, which often go unnoticed. This is problematic as it can lead to significant bias, large variance and invalid inference. We will illustrate this problem for different estimators, e.g. imputation and weighting methods, using real case studies. Moreover, we show that estimands used to handle intercurrent events are often too ambitious and cannot be inferred from the data without relying on very strong assumptions. In the second part of the talk, we will discuss different solutions in terms of estimands, estimators and trial design.

Thursday, 07/Sept/2023 11:20am - 11:40am
ID: 275 / S64: 3
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), Estimands and causal inference
Keywords: estimand, longitudinal outcomes, intercurrent event, data-generating model, simulation study

Data-generating models of longitudinal continuous outcomes and intercurrent events to evaluate estimands

Marian Mitroiu[1,2], Steven Teerenstra[1,3], Katrien Oude Rengerink[1,2], Frank Petavy[4], Kit Roes[1,3]
[1]Methodology Working Group, Medicines Evaluation Board, The Netherlands; [2]Clinical Trial Methodology Department, Biostatistics and Research Support, Julius Center for Health Sciences and Primary Care, Biostatistics and Research Support, University Medical Center Utrecht, Utrecht University, The Netherlands; [3]Department for Health Evidence, section Biostatistics, Radboud University Medical Center, The Netherlands; [4]Data Analytics and Methods Taskforce, European Medicines Agency, The Netherlands; marian.mitroiu@icloud.com

Introduction/Background: We aimed to develop and evaluate data-generating models to jointly simulate outcomes and intercurrent events for randomised clinical trials to enable the assessment of properties of estimands.

Methods: We propose four data-generating models for the joint distribution of longitudinal continuous clinical outcomes and intercurrent events under the scenario where they are observable: a selection model, a pattern-mixture mixed model, a shared-parameter model and a joint model of longitudinally observed clinical outcomes and a survival model for intercurrent events. We present a case study in a short-term depression trial with repeated measurements of continuous outcomes and two types of intercurrent events, and compare the four proposed data-generating models.

Results: In our case study, we found that all four data-generating models can simulate different types of intercurrent events, their timing, and their associated longitudinal outcomes. These can be used to match envisaged patterns of intercurrent events and outcomes informed by prior available clinical trial data. For a given intercurrent event, the Shared-Parameter and Joint Models tend to associate more similar longitudinal profiles (because of shared latent random effects), while the Selection Model and Pattern-Mixture Model could allow more variation in associated profiles.

Conclusion: All four proposed data-generating models can be used to evaluate different estimands and to investigate their properties in-depth in the design stage. Thereby they are useful tools for the selection of estimands a priori.

Wednesday, 06/Sept/2023 8:50am - 9:10am
ID: 337 / **S45: 2**
Presentation Submissions - Regular Session (Default)
Topic of Submission: High dimensional data, genetic and x-omics data, Statistical hypothesis testing (covariate adjustment, nonparametric methods, multiple comparisons, …)
Keywords: Multiple contrast tests, hypothesis testing, type-I error rate control, high-dimensional designs, bootstrap, small sample sizes

Maximum Test Method for the Wilcoxon-Mann-Whitney Test in High-Dimensional Designs

Lukas Mödl, Frank Konietschke
Institut für Biometrie und Klinische Epidemiologie -- Chartié Berlin, Germany; lukas.moedl@cahrite.de

The statistical comparison of two multivariate samples is a frequent task, e.g. in biomarker analysis. Parametric and nonparametric multivariate analysis of variance (MANOVA) procedures are well established procedures for the analysis of such data. Which method to use depends on the scales of the endpoints and whether the assumption of a parametric multivariate distribution is meaningful. However, in case of a significant outcome, MANOVA methods can only provide the information that the treatments (conditions) differ in any of endpoints; they cannot locate the guilty endpoint(s). Multiple contrast tests in terms as maximum tests on the contrary provide local test results and thus the information of interest.

The maximum test method controls the error rate by comparing the value of the largest contrast in magnitude to the (1-α)-equicoordinate quantile of the joint distribution of all considered contrasts. The advantage of this approach over existing and commonly used methods that control the multiple type-I error rate, such as Bonferroni, Holm, or Hochberg, is that it is appealingly simple, yet has sufficient power to detect a significant difference in high-dimensional designs, and does not make strong assumptions (such as MTP2) about the joint distribution of test statistics. Furthermore, the computation of simultaneous confidence intervals is possible. The challenge, however, is that the joint distribution of the test statistics used must be known in order to implement the method.

In this talk, we develop a simultaneous maximum Wilcoxon-Mann-Whitney test for the analysis of multivariate data in two independent samples. We hereby consider both the cases of low-and high-dimensional designs. We derive the (asymptotic) joint distribution of the test statistic and propose different bootstrap approximations for small sample sizes. We investigate their quality within extensive simulation studies. It turns out that the methods control the multiple type-I error rate well, even in high-dimensional designs with small sample sizes. A real data set illustrates the application.

Thursday, 07/Sept/2023 9:30am - 9:50am
ID: 251 / **S62: 4**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Epidemic short-term forecasting in real time
Keywords: covid 19 forecast

Strong effect of testing in containing Covid-19

Jan Mohring, Neele Leithäuser, Jaroslaw Wlazlo, Marvin Schulte, Johanna Münch, Maximilian Pilz
Fraunhofer ITWM, Germany; jan.mohring@itwm.fraunhofer.de

With the outbreak of the Covid-19 epidemic in Germany, a spread model was developed at Fraunhofer ITWM, which is successfully used for short-term forecasts. The results have entered the German-Polish and European Forecast Hub and still form the basis for advising the state of Rhineland-Palatinate. We start the talk summarizing briefly the integral equation-based model which fits contact and detection rates to reproduce counted numbers of cases and deaths. In a first analysis we consider the situation in Germany in spring 2021. In this phase of the pandemic, data was still available at high quality and daily resolution for different regions and age groups. Moreover, protection could easily be described by three states: susceptible, vaccinated, or recovered. These facts made it possible to fit even the characteristics of the virus, like incubation time and infectious period, without consulting literature. In contrast to many other groups at that time, we fitted explicitly the detection rate and the offset between infection and detection. This enabled us to reconstruct the actual effect of post-test isolation. Comparing the different trajectories of reporting data around the staggered Easter holidays in selected German states, we bring strong evidence that, in spring 2021, post-test isolation made a stronger contribution to the containment of Covid-19 than vaccination or contact restrictions. For this purpose, we are fitting contact rates, which change with contact restrictions, and detection rates, which change with the testing regime. Freezing two of the three rates for restrictions, testing, and vaccination at their values before Easter 2021, and continuing the third one as fitted, we can estimate the effect of each individual measure. It turns out that, in particular, tests at schools have played an important role. In the remaing part of the talk we will discuss recent improvements to our code dealing with transitions between variants and incorporation of waste water measurements.

Monday, 04/Sept/2023 12:00pm - 12:20pm
ID: 211 / S7: 4
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), Statistical hypothesis testing (covariate adjustment, nonparametric methods, multiple comparisons, …), Preclinical drug development, safety and toxicology
Keywords: alert concentrations, concentration-response modelling, gene expression data, parametric bootstrap, relevant hypotheses

Identifying alert concentrations using a model-based bootstrap approach

Kathrin Möllenhoff[1], Kirsten Schorning[2], Franziska Kappenberg[2]
[1]Mathematical Institute, Heinrich Heine University, Düsseldorf, Germany; [2]Department of Statistics, TU Dortmund University, Dortmund, Germany; kathrin.moellenhoff@hhu.de

The determination of alert concentrations, where a pre-specified threshold of the response variable is exceeded, is an important goal of concentration–response studies. The traditional approach is based on investigating the measured concentrations and attaining statistical significance of the alert concentration by using a multiple t-test procedure.

In this talk, we propose a new model-based method to identify alert concentrations, based on first fitting a concentration–response curve and second constructing a simultaneous confidence band for the difference of the response of a concentration compared to the control. In order to obtain these confidence bands, we use a bootstrap approach which can be applied to any functional form of the concentration–response curve. This particularly offers the possibility to investigate also those situations where the concentration–response relationship is not monotone and, moreover, allows to detect alerts at concentrations which were not measured during the study, providing a highly flexible framework for the problem at hand.

We demonstrate the validity of the method by means of a simulation study and present an application to a real dataset investigating the effect of different concentrations of the compound VPA on the development of hESC to neuroectoderm.

Wednesday, 06/Sept/2023 11:00am - 11:20am
ID: 450 / **S51: 2**
Presentation Submissions - Featured Session
Featured Sessions: IBS-DR/ROeS Award session
Keywords: blinded sample size reestimation, event-driven designs, extrapolation, flexible parametric models, splines

Blinded sample size reestimation in clinical trials with time-to-event outcomes based on flexible parametric models

Tim Mori[1], Sho Komukai[2], Satoshi Hattori[2], Tim Friede[3,4]
[1]German Diabetes Center (DDZ), Germany; [2]Osaka University, Japan; [3]University Medical Center Göttingen, Germany; [4]DZHK (German Center for Cardiovascular Research), Germany; tim.mori@ddz.de

For event-driven designs the objective is to complete a clinical trial within a given time frame. Blinded sample size reestimation (BSSR) methods use non-comparative blinded interim trial data to adjust the sample size if the planning assumptions are wrong. In clinical trials with time-to-event outcomes the estimated survival function based on the interim data needs to be extrapolated for the purpose of BSSR. The current practice is to fit standard parametric models (e.g. exponential or Weibull models), which may however not always be suitable. For example, some real life datasets exhibit complex hazard functions that cannot be captured by simple parametric models. The aim of the current article was to propose a flexible parametric approach for BSSR in clinical trials with time-to-event outcomes and to compare it to existing parametric approaches. Specifically, we propose to carry out the extrapolation based on the Royston-Parmar spline model. We carried out a simulation study based on exponential, Weibull and Gompertz distributed data and considered a practical application of spline BSSR to a Secondary Progressive Multiple Sclerosis (SPMS) trial. In our simulation study we found that a 1-knot spline BSSR was unbiased in the exponential and Weibull setting and performed best in the Gompertz misspecification scenario. In our case study we found spline BSSR to perform well and to outperform the Weibull BSSR. Overall, if planning assumptions are wrong this more robust spline BSSR could help event-driven designs to more accurately adjust recruitment numbers and to finish on time.

Monday, 04/Sept/2023 2:00pm - 2:20pm
ID: 486 / S8: 1
Presentation Submissions - Featured Session
Featured Sessions: Statistics in Practice: Simulation studies as a tool to evaluate and compare the properties of statistical methods – an overview

Simulation studies as a tool to assess and compare the properties of statistical methods – an overview

Tim Morris, **Brennan C. Kahan**
MRC Clinical Trials Unit at UCL; tim.morris@ucl.ac.uk, b.kahan@ucl.ac.uk

Simulation studies are a key tool for studying the properties of statistical methods. As such, they contribute to the evidence base that supports the choice of methods in practice. This workshop will provide an overview of simulation studies for those who might use them. The workshop will be split into four sessions, providing a whistle stop tour of the planning, coding, analysis and reporting of simulation studies.

1 Planning

Simulation studies need to be carefully planned. This session will outline the ADEMP structure for planning simulation studies, which involves defining Aims, Data-generating mechanisms, Estimands, Methods of analysis and Performance measures. We will discuss issues and subtleties that should be considered for each of the steps, and introduce standard terminology.

2 Coding

All simulation studies require coding. This can be extremely simple but is frequently complex, and it is easy to make mistakes. In this session, we will cover how to write 'defensive' code for a simulation study that reduces the chance of making mistakes that produce misleading results. As such, we will focus on the concepts and ideas but not actual pieces of code.

3 Analysis

Once a simulation study has been run, it needs to be analysed. Analysis should begin by checking the data (e.g. for missing values). This session will focus on performance measures: the metrics by which we judge results. We will define some common performance measures and describe how they are estimated, emphasising the importance of simulation uncertainty (Monte Carlo error). The session will end by considering how to choose the sample size for a simulation study.

4 Writing up and reporting

Not every simulation study is intended for publication. However, if a simulation study is to be published and its results trusted, it must be to be clearly understood by others. This session will describe some principles for tabular and graphical displays of simulation results, and consider some examples from the literature. We will make suggestions for what should be reported and, of course, advocate open sharing of code

Some key publications

[1] T. P. Morris, I. R. White, and M. J. Crowther, "Using simulation studies to evaluate statistical methods," Statistics in Medicine, vol. 38, no. 11, pp. 2074–2102, 2019. doi: 10.1002/sim.8086.

[2] A.-L. Boulesteix, H. Binder, M. Abrahamowicz, and W. Sauerbrei, for the Simulation Panel of the STRATOS Initiative, "On the necessity and design of studies comparing statistical methods," Biometrical Journal, vol. 60, no. 1, pp. 216–218, Nov. 2017. doi: 10.1002/bimj.201700129. [Online]. Available: https://doi.org/10.1002/ bimj.201700129.

[3] G. Heinze, A.-L. Boulesteix, M. Kammer, T. P. Morris, and I. R. White, "Phases of methodological research in biostatistics—building the evidence base for new methods," Biometrical Journal, p. 2 200 222, 2023. doi: 10.1002 / bjm. 202200222. [Online]. Available: https://doi.org/10.1002/bimj.202200222.

[4] I. R. White, T. M. Pham, M. Quartagno, and T. P. Morris, "How to check a simulation study," 2023. doi: 10.31219/osf.io/cbr72. [Online]. Available: https://doi. org/10.31219/osf.io/cbr72.

Tuesday, 05/Sept/2023 4:50pm - 5:10pm
ID: 124 / **S38: 2**
Presentation Submissions - Invited Session
Invited Sessions: From multivariate to high-dimensional and functional data
Keywords: functional data, multiple testing, resampling, heteroscedasticity

MultiFANOVA: Multiple Contrast Tests for Functional Data

Merle Munko[1], Marc Ditzhaus[1], Markus Pauly[2], Łukasz Smaga[3]
[1]Otto-von-Guericke University, Magdeburg, Germany; [2]TU Dortmund University, Dortmund, Germany; [3]Adam Mickiewicz University, Poznań, Poland; merle.munko@ovgu.de

In various scientific fields, functional data can be observed more and more frequently. This includes audiology, biology, ergonomy, meteorology, growth studies and environmentology (Zhang, 2013, p. 2), to name just a few examples.

Several tools for the analysis of functional data, including tests for complex factorial designs and functional ANOVA problems, have already been studied (e.g. Zhang, 2013). However, they mainly rely on homoscedasticity assumptions, which are often not justifiable in practice. Moreover, all these strategies are designed for global null hypotheses testing, e.g. for main and interaction effects, and do not directly allow a more in-depth analysis by testing several null hypotheses simultaneously.

To address the first problem, we obtain a test function that takes the heteroscedasticity of the functional data into account. Integrating over the test function yields a test statistic for general null hypotheses in factorial designs, which has a rather complicated limit null distribution. Therefore, we propose a resampling procedure to approximate the null distribution. In a next step, we explain how to use the described testing strategy and resample scheme to infer several local null hypotheses simultaneously. Hereby, we incorporate the asymptotic exact dependency structure between the local test statistics to avoid a significant power loss. The resulting multiple testing procedure is consonant and coherent as defined in Gabriel (1969). Moreover, the proposed multiple contrast tests control the level of significance of the global test as well as the family-wise type I error rate. The small sample performances of the proposed global and multiple testing procedures are analyzed in extensive simulations and finally illustrated by analyzing a real data example.

References:

1. K.R. Gabriel (1969). Simultaneous test procedures–some theory of multiple comparisons. The Annals of Mathematical Statistics, 40(1):224-250.
2. J.-T. Zhang (2013). Analysis of variance for Functional Data. Chapman & Hall/CRC.

Tuesday, 05/Sept/2023 4:30pm - 4:50pm
ID: 191 / **S37: 2**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Causal estimands for time to event data
Keywords: estimands, time-to-event outcomes, clinical question

Treatment effect measures in clinical trials with time-to-event outcomes: it is time to apply estimand thinking

Tobias Mütze[1], **Vivian Lanius**[2]
[1]Novartis Pharma AG, Basel, Switzerland; [2]Bayer AG, Wuppertal, Germany; tobias.muetze@novartis.com, vivian.lanius@bayer.com

The ICH E9(R1) addendum on estimands and sensitivity analysis in clinical trials calls for clarity and precision when describing the clinical question of interest. It defines the estimand as a population-level summary of “what the outcomes would be in the same patients under different treatment conditions being compared”. Thus, while not explicitly using the term “causal”, both the framework and language used in ICH E9(R1) are aligned with causal reasoning.

In randomized clinical trials with a time to event endpoint, the hazard ratio is still the most common effect measure. Post-randomization (i.e., intercurrent) events are often addressed through censoring without explicitly discussing or stating the underlying clinical question of interest. Alternative summary measures, especially on a probability scale or time scale, are rarely considered in clinical trials despite being seemingly easier to interpret and potentially more meaningful to patients and practitioners.

In this talk we will present the status of ongoing discussions among a working group of statisticians from different pharmaceutical companies on estimands for clinical trials with time-to-event data. In detail, we will discuss what key clinically meaningful questions of interest are when measuring the effect of an intervention through a time-to-event endpoint. We will reflect on the interpretation of various summary measures, the role of causality when defining an estimand in a clinical trial, and on how the choice of the estimand affects the design of a trial with a time-to-event endpoint. We will also elaborate on the practicalities of summarizing the effect of treatment through a single number in a time to event setting and discuss separating testing and estimation.

Wednesday, 06/Sept/2023 10:40am - 11:00am
ID: 452 / **S56: 1**
Presentation Submissions - Featured Session
Featured Sessions: Best practices for Data Monitoring Committees and how to get there
Keywords: DMC

Interactive workshop on communicating data to Data Monitoring Committees

Tobias Mütze, **David Lawrence**
Novartis Pharma AG, Switzerland; tobias.muetze@novartis.com, David.Lawrence@novartis.com

Emerging data is presented to a Data Monitoring Committee (DMC) prior to the data review meeting through a DMC report. This interactive session aims to illustrate difficulties a DMC may have with a poorly designed report, to share experiences of working with DMCs or being members of DMCs and discuss suggestions for improvements to DMC reports including newer technologies such as apps.

Wednesday, 06/Sept/2023 9:50am - 10:10am
ID: 437 / **S45: 5**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), Statistical hypothesis testing (covariate adjustment, nonparametric methods, multiple comparisons, …)
Keywords: Time-series, FoSR, Multiple testing

A novel approach to Function-on-Scalar Regression (FoSR) for the analysis of Periodic Time-Series

Konrad Neumann
Charité, Germany; konrad.neumann@charite.de

Analysis of periodic time-series plays an important role in research dealing with data from wearables such as smart watches or accelerometer devices. Function-on-scalar regression (FoSR) is a popular method of analysing such data ([1] and [2]). FoSR is a family of multivariate regression models that describe the association of covariates with time-series as a response. In the talk, a novel approach to FoSR will be presented. It follows the ideas of classical least squares analysis of the general linear model. In contrast to the classical approach, the components of the response vector may now be points in an arbitrary Hilbert space H, such as the L_2 space. The coefficient functions and their least squares estimates are then points in a predefined finite dimensional subspace H' of H. Furthermore, choosing the H' carefully leads to a two-step multiple testing procedure that bounds the familywise error rate of first kind. For this version of FoSR only little assumptions must be made. Furthermore, this classical approach leads to explicit formulae of the coefficient functions and of the test statistics. An example from [3] will illustrate the method.

References

1. Goldsmith, J., Liu, X., Jacobson, J. S. & Rundle, A. New Insights into Activity Patterns in Children, Found Using Functional Data Analyses. *Med. Sci. Sports Exerc.* **48,** 1723–1729
2. Xiao, L. *et al.* Quantifying the lifetime circadian rhythm of physical activity: A covariate-dependent functional approach. *Biostatistics* **16,** 352–367 (2015).
3. Rackoll, T., Neumann, K., Passmann S., Grittner, U., Külzow, N., Ladenbauer, J., Floel, A. Applying time series analyses on continuous accelerometry data- A clinical example in older adults with and without cognitive impairment. PloS One 16(5) (2021).

Tuesday, 05/Sept/2023 4:10pm - 4:30pm
ID: 348 / **S42: 1**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Free Contributions
Keywords: Simulation; Real Data; Comparison studies; Neutrality

Towards more practically relevant method comparison studies by generating simulations based on a sample of real data sets

Christina Nießl[1,2], Maria Thurow[3], Ina Dormuth[3], Markus Pauly[3,5], Marc Ditzhaus[4], Anne-Laure Boulesteix[1]
[1]LMU Munich, Germany; [2]Munich Center for Machine Learning (MCML); [3]TU Dortmund University; [4]Otto-von-Guericke-University Magdeburg; [5]UA Ruhr, Research Center Trustworthy Data Science and Security; cniessl@ibe.med.uni-muenchen.de

Simulation studies are an essential approach for comparing statistical methods. Two of the key advantages that set them apart from benchmark studies based on real data are (1) the availability of the ground truth and (2) the wide range of parameters that can be explored. However, these features come at a price: Simulation studies are often criticized for being too simplistic and not reflecting reality. Moreover, the infinite parameter space presents researchers with the often difficult decision of choosing realistic and informative parameter values. This process is also prone to the selective reporting of parameter values that lead to favorable results (e.g., a good performance for a specific method), a questionable research practice that threatens the neutrality of simulation studies.

To overcome these drawbacks, several approaches have been proposed to design simulation studies based on real data, for example setting key parameters (e.g., sample sizes, means, variances, or correlations) according to real data examples. However, the number of underlying data sets is usually restricted to one or two, and it is often not clear how these data sets were selected.

In this work, we present the idea of systematically basing simulations on a whole sample of real data sets that were selected according to pre-specified inclusion criteria as a means to obtain comprehensive and practically relevant results. We illustrate this approach using two examples. For the first example, we simulate data reflecting two-arm trials with ordinal endpoints. Here, the parameter of interest is the distribution of the ordinal endpoint in the two treatment groups. We set this parameter by sampling from all articles that were published in selected issues of the New England Journal of Medicine and that analyzed two-arm trials with ordinal endpoints.

For the second example, we consider the comparison of differential gene expression methods that aim to identify genes with differences in their expression levels between two conditions. In this more complex simulation, there are several parameters to be specified, such as the mean expression level or dispersion of each gene. In this application, we specify our sample as all cancer data sets provided in the The Cancer Genome Atlas (TCGA) data base.

For both examples, the results based on the sampled simulation parameters differ from the results of user-specified parameters and parameters based on a single data set, hence suggesting the potential of "simulation-sampling" as a useful complement to standard simulation approaches.

Tuesday, 05/Sept/2023 4:30pm - 4:50pm
ID: 121 / **S36: 2**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical hypothesis testing (covariate adjustment, nonparametric methods, multiple comparisons, …)
Keywords: Brunner-Munzel test, error spending, group sequential methods, nonparametric relative effect, Wilcoxon-Mann-Whitney test

Group sequential methods for the Mann-Whitney parameter

Claus Nowak[1], Tobias Mütze[1], Frank Konietschke[2]
[1]Novartis Pharma AG, Switzerland; [2]Charité - Universitätsmedizin Berlin; claus.nowak@novartis.com

Late phase clinical trials are occasionally planned with one or more interim analyses to allow for early termination or adaptation of the study. While extensive theory has been developed for the analysis of ordered categorical data in terms of the Wilcoxon-Mann-Whitney test, there has been comparatively little discussion in the group sequential literature on how to provide repeated confidence intervals and simple power formulas to ease sample size determination. Dealing more broadly with the nonparametric Behrens-Fisher problem, we focus on the comparison of two parallel treatment arms and show that the Wilcoxon-Mann-Whitney test, the Brunner-Munzel test, as well as a test procedure based on the log win odds, a modification of the win ratio, asymptotically follow the canonical joint distribution. In addition to developing power formulas based on these results, simulations confirm the adequacy of the proposed methods for a range of scenarios. Lastly, we apply our methodology to the FREEDOMS clinical trial (ClinicalTrials.gov Identifier: NCT00289978) in patients with relapse-remitting multiple sclerosis.

doi:https://doi.org/10.1177/09622802221107103

Thursday, 07/Sept/2023 9:50am - 10:10am
ID: 152 / **S62: 5**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Epidemic short-term forecasting in real time
Keywords: agent-based model (ABM), epidemic dynamics, epidemiology, immunity, COVID-19

Multi-step immunity mechanism in ICM UW epidemic agent-based model (PDYN 1.5)

Jędrzej M. Nowosielski[1], Grzegorz Dudziuk[1], Magdalena Gruziel-Słomka[1], Karol Niedzielewski[1], Maciej Radwan[1], Antoni Moszyński[1], Jakub Zieliński[1], Rafał P. Bartczuk[1], Dominik Bogucki[1], Filip Dreger[1], Łukasz Górski[1], Jędrzej Haman[1], Artur Kaczorek[1], Jan Kisielewski[2], Bartosz Krupa[1], Marcin Semeniuk[1], Franciszek Rakowski[1]
[1]Interdisciplinary Centre for Mathematical and Computational Modelling University of Warsaw (ICM UW)), Warsaw, Poland;
[2]Faculty of Physics, University of Bialystok, Bialystok, Poland; j.nowosielski@icm.edu.pl

In PDYN 1.5—the epidemiological agent-based model developed at ICM UW—for each agent the probability of contracting infection is calculated daily. Once the agent gets infected, the course of infection consists of a chain of subsequent disease states that the agent goes through. These states represent the severity of symptoms such asymptomatic, mild symptomatic, requiring hospitalisation or requiring ICU treatment. The probabilities of transition from each state to other states determine the total probabilities of each particular course of the whole infection.

The probability of infection as well as transition probabilities between states may be reduced due to the agent's immunity gained from vaccination or past infection within the frame of a multi-step immunity mechanism. The proposed immunity mechanism captures well-known characteristic of COVID-19 vaccines, i.e., much better efficacy against the severe outcomes of the infection than against the infection itself. The mechanism allows for differentiation of the probabilities of transition to particular severe disease states in the vaccinated and non-vaccinated groups of agents.

Using PDYN 1.5 calibrated to the Polish epidemiological data, we were able to show that our model along with the above mentioned immunity mechanisms is able to capture the reduction of the risk of hospitalisation relative to the risk of the infection itself resulting from vaccination or past infection. Moreover, our method proved its validity in the predictions of delta and omicron VoCs waves in 2021/2022 winter season that were published in Covid-19 European forecast hub. Additionally, to gain further insight into our model of immunity, we carried out an in-silico epidemiological study in which we evaluate the vaccine efficacy (VE) in our simulations and compare it to the real-life data.

In this talk, we would like to present these results as well as explain the mentioned immunity mechanisms implemented in PDYN 1.5 in more detail.

Monday, 04/Sept/2023 12:00pm - 12:20pm
ID: 383 / **S3: 4**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), High dimensional data, genetic and x-omics data, Real world data and evidence, Statistical hypothesis testing (covariate adjustment, nonparametric methods, multiple comparisons, …)
Keywords: polygenic score, asthma, prediction performance

Investigating different numbers of variants in polygenic scores using the ALLIANCE cohort

Lisa-Marie Nuxoll[1], Lea Louisa Kronziel[1,2], Inke R. König[1,2]
[1]Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck, Germany; [2]Airway Research Centre North (ARCN), Member of the German Centre for Lung Research (DZL), Lübeck, Germany; l.nuxoll@uni-luebeck.de

A polygenic score (PGS) can be used to estimate an individual's genetic liability to a trait or disease. For this score, the individual's genotype information is weighted with results from a genome-wide association study (GWAS) to calculate an individual score for the observed trait or disease. As sample sizes of GWAS increase, PGS may become more powerful in the near future and will be valuable in personalized medicine. However, the value and usefulness of PGS depend on method development to construct PGS, proper use of PGS in analysis and appropriate interpretation of results.

To date, new PGS for various traits are constantly being developed and published. Importantly, new PGS may be developed for a particular trait (e.g., lung function), even though PGS are already available for that trait. These PGS for the same trait might then differ in terms of the selection of variants and/or in the number of variants they contain. The number of variants in a PGS is likely to affect the strength of association with the trait and may also play a role in subsequent clinical applications. Thus, a larger number of variants in a PGS may provide a more precise prediction, but, at the same time, may have the disadvantage that many required variants may not be present in the target dataset, making replication challenging. In addition, a larger number of variants in a PGS may lead to overfitting. This trade-off has not been analyzed systematically.

Therefore, our study investigates the effect of the number of variants in previously published PGS on prediction performance for the lung function traits FEV1/FVC [1,2] and FEV1 [2]. The PGS were calculated for participants of a German pediatric asthma cohort (ALLIANCE) [3] including 526 children with asthma and 249 children without asthma. The considered PGS use 279 variants [1], 1,713,430 and 1,232,916 variants [2]. After calculating the PGS for ALLIANCE participants, the distributions of the PGS were examined and various association analyses were performed. Additional clinical variables were also considered, as models with clinical and genetic information can provide higher accuracy than models containing only genetic information.

The calculation of the PGS published by Moll et al. was found to be problematic because they contain many variants that are not present in the available genetic data of the ALLIANCE cohort. Therefore, an extensive proxy search had to be performed, which may lead to less accuracy and more bias. Subsequent association analyses showed no associations between the observed PGS by Shrine et al. and the asthma phenotypes in the ALLIANCE cohort. However, since the number of variants in the considered PGS is very different, no conclusion about overfitting can be drawn.

Literature:

1. Shrine, N., Guyatt, A.L., Erzurumluoglu, A.M. et al. 2019 Nat Genet 51:481–93; doi: https://doi.org/10.1038/s41588-018-0321-7
2. Moll, M., Sakornsakolpat, P. et al. 2020 Lancet Respir Med 8: 696–708; doi: https://doi.org/10.1016/S2213-2600(20)30101-6
3. Fuchs, O., Bahmer, T., Weckmann, M. et al. 2018 BMC Pulm Med 18:140; doi: https://doi.org/10.1186/s12890-018-0705-6

Monday, 04/Sept/2023 2:00pm - 2:20pm
ID: 335 / **S10: 1**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Personalized health care, Real world data and evidence
Keywords: Real world data, Semantics, FAIR, National Infrastructure, clinical routine data

Swiss Personalized Health Network from clinical routine data to FAIR data for research

Sabine Oesterle, Katrin Crameri
SIB Swiss Institute of Bioinformatics, Switzerland; sabine.oesterle@sib.swiss

The Swiss Personalized Health Network (SPHN) has taken an innovative approach to address the challenge of utilizing real-world data for research purposes in a scalable and sustainable way. By developing a national framework and tool stack for the semantic representation of health data in a knowledge graph, SPHN has created a solution that enables the sharing and integration of various types of health-related data from different sources. By using semantic standards like SNOMED CT, the knowledge of these ontologies can be used to enrich the semantics of the individual data elements. This is a significant step forward in the transition from data to knowledge.

The SPHN approach has been implemented in all Swiss university hospitals, allowing for the semantic interoperability of data between sites and facilitating the linking of clinical routine data with other data sources, such as omics data or data from clinical research studies. The framework complies with the FAIR principles, making the data findable, accessible, interoperable, and reusable.

By implementing this framework, Switzerland is now poised to leverage its rich clinical routine data for research purposes, enabling researchers to answer important questions related to personalized medicine and beyond. This will lead to significant advances in secondary use of data, particularly in medical research, and ultimately better health outcomes for patients. The SPHN framework can serve as a model for other countries seeking to harness the power of real-world data to accelerate scientific discovery and innovation.

Tuesday, 05/Sept/2023 3:00pm - 3:20pm
ID: 273 / S34: 4
Presentation Submissions - Regular Session (Default)
Topic of Submission: Meta-Analysis and Systematic Reviews
Keywords: evidence synthesis, prognostic model, multiple sclerosis, machine learning

Prognostic models for disease progression in people with multiple sclerosis – a systematic review and assessment of current methodological challenges

Begum Irmak On*[1], Kelly A. Reeve*[2], Joachim Havla[3], Jacob Burns[1], Martina Gosteli[4], Ulrich Mansmann[1], Ulrike Held[2]
[1]Biometrics and Bioinformatics, Ludwig-Maximilians University, Germany; [2]Department Biostatistics, University of Zurich, Switzerland; [3]Institute of Clinical Neuroimmunology, LMU Hospital, Ludwig-Maximilians University Munich, Germany; [4]University Library, University of Zurich, Switzerland; * equal contribution; ionseker@ibe.med.uni-muenchen.de

Systematic reviews of prognostic models have revealed poor reporting quality, that studies suffer from high risk of bias, and lack of external validation studies [1, 2]. This is true also for studies published after the release of the TRIPOD reporting guideline [3]. In a Cochrane systematic review, we aimed to identify and summarize multivariable prognostic models for quantifying the risk of clinical disease progression, worsening, and activity in adults with multiple sclerosis (MS)[4]. Relevant databases were searched from January 1996, when an important tutorial on multivariable prognostic models was published [5], to July 2021. Studies evaluating performance (i.e., validation studies) were also included. More than 13,000 records were identified through the database search, and of these 57 studies were identified, reporting on 75 model developments. Of these, 35 models were developed using traditional statistical methods, and the remaining using machine learning (ML) methods. Only two of these models were evaluated externally multiple times. None of the validations were performed by researchers independent from those that developed the model. Over half (52%) of the developed models were not accompanied by model coefficients, tools, or instructions, which hinders their application, independent validation, or reproduction. All but one of the model developments or validations was rated as having high overall risk of bias. The main reason for this was the statistical methods used for the development or evaluation of prognostic models. Over time, we observed an increase in the percent of participants on treatment, diversification of the diagnostic criteria used, an increase in consideration of biomarkers or treatment as predictors, and increased use of ML methods, with the first being published in 2009. Major reporting deficiencies were observed, these reporting deficiencies were more pronounced in the studies using ML. We conclude that current evidence is not sufficient for recommending the use of any of the published prognostic models for people with MS. The MS prognostic research community should adhere to the current reporting and methodological guidelines and conduct many more state-of-the-art external validation studies for the existing or newly developed models. Gaps in methodological guidance were identified regarding the assessment of models developed using complex ML methods.

References:

1. Wynants, L., et al., *Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal.* BMJ, 2020. **369**: p. m1328.
2. Van Grootven, B., et al., *Prediction models for hospital readmissions in patients with heart disease: a systematic review and meta-analysis.* BMJ Open, 2021. **11**(8): p. e047576.
3. Collins, G.S., et al., *Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement.* BMJ, 2015. **350**: p. g7594.
4. On Seker, B.I., Reeve, K., Havla, J., Burns, J., Gosteli, M., Lutterotti, A., Schippling, S., Mansmann, U., Held, U., *Prognostic models for predicting clinical disease progression, worsening and activity in people with multiple sclerosis* Cochrane Library Protocol, 2020.
5. Harrell, F.E., Jr., K.L. Lee, and D.B. Mark, *Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors.* Stat Med, 1996. **15**(4): p. 361-87.

Tuesday, 05/Sept/2023 11:40am - 12:00pm
ID: 423 / **S28: 3**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Meta-Analysis and Systematic Reviews
Keywords: Clinical trials, COVID-19, between-trial heterogeneity, hierarchical models, random-effects, subgroups

Random-effects meta-analysis of subgroup specific effects and treatment-by-subgroup interactions

Renato Valladares Panaro, Christian Röver, Tim Friede
University Medical Center Göttingen, Germany; renato.panaro@med.uni-goettingen.de

Random-effects meta-analysis of subgroup specific effects and treatment-by-subgroup interactions

Renato Panaro, Christian Röver and Tim Friede

Contrast-based meta-analysis investigates whether a treatment is more effective than a reference across a set of controlled trials. In addition to the main effect analysis, studies commonly report effect estimates by subgroups based on some baseline characteristics (e.g., sex or age groups). This way, meta-analysis methods considering subgroup information may be able to identify whether or which patient subgroups benefit most from an intervention by explicitly accounting for treatment-by-subgroup interaction effects.

Several meta-analyses have been proposed in this context, modeling main effects and interactions jointly or separately (Godolphin et al, 2020; van Houwelingen et al, 2002). When this is done separately, standard methods for meta-analysis can be applied. However, the results might not be consistent across the analyses. To avoid such inconsistencies, recently an estimation matching strategy has been proposed to first analyze subgroup contrasts and subsequently derive reference subgroup effects from the interaction residuals (Godolphin et. al 2020). However, estimate matching has some disadvantages when it comes to subgroup effect estimation, especially when not all subgroups are represented in all trials, or when heterogeneity in treatment-by-subgroup interactions is considered. This work investigates and compares the different subgroup estimators regarding their statistical properties and operational performance in simulation studies. The methods are motivated and illustrated using recent treatment trials in COVID-19 (WHO 2020, 2021).

References:

1. Godolphin, PJ, White, IR, Tierney, JF, Fisher, DJ. Estimating interactions and subgroup-specific treatment effects in meta-analysis without aggregation bias: A within-trial framework. *Res Syn Meth*. 2023; 14(1): 68- 78. doi:10.1002/jrsm.1590
2. The WHO Rapid Evidence Appraisal for COVID-19 Therapies (REACT) Working Group. Association Between Administration of Systemic Corticosteroids and Mortality Among Critically Ill Patients With COVID-19: A Meta-analysis. *JAMA*. 2020;324(13):1330–1341. doi:10.1001/jama.2020.17023
3. The WHO Rapid Evidence Appraisal for COVID-19 Therapies (REACT) Working Group. Association Between Administration of IL-6 Antagonists and Mortality Among Patients Hospitalized for COVID-19: A Meta-analysis. *JAMA*. 2021;326(6):499–518. doi:10.1001/jama.2021.11330
4. van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. Stat Med. 2002 Feb 28;21(4):589-624. doi: 10.1002/sim.1040.

Monday, 04/Sept/2023 11:00am - 11:20am
ID: 120 / **S1: 1**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Neutral comparison studies in methodological research
Keywords: benchmarking studies, Monte Carlo experiments, overoptimism, reproducibility, transparency

Pitfalls and Potentials in Simulation Studies

Samuel Pawel, Lucas Kook, Kelly Reeve
University of Zurich, Switzerland; samuel.pawel@uzh.ch

Comparative simulation studies are workhorse tools for benchmarking statistical methods. As with other empirical studies, the success of simulation studies hinges on the quality of their design, execution and reporting. If not conducted carefully and transparently, their conclusions may be misleading. In this paper we discuss various questionable research practices which may impact the validity of simulation studies, some of which cannot be detected or prevented by the current publication process in statistics journals. To illustrate our point, we invent a novel prediction method with no expected performance gain and benchmark it in a pre-registered comparative simulation study. We show how easy it is to make the method appear superior over well-established competitor methods if questionable research practices are employed. Finally, we provide concrete suggestions for researchers, reviewers and other academic stakeholders for improving the methodological quality of comparative simulation studies, such as pre-registering simulation protocols, incentivizing neutral simulation studies and code and data sharing.

Wednesday, 06/Sept/2023 9:30am - 9:50am
ID: 387 / S45: 4
Presentation Submissions - Regular Session (Default)
Topic of Submission: Machine Learning and Data Science, High dimensional data, genetic and x-omics data, Statistical hypothesis testing (covariate adjustment, nonparametric methods, multiple comparisons, …)

Mind your zeros: accurate p-value approximation in permutation testing with applications in microbiome data analysis

Stefanie Peschel[1,2], Martin Depner[3], Erika von Mutius[3,4,5,6], Anne-Laure Boulesteix[2,7], Christian L Müller[1,2,8,9]
[1]Department of Statistics, LMU München, Munich, Germany; [2]Munich Center for Machine Learning, Munich, Germany; [3]Institute of Asthma and Allergy Prevention, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany; [4]Department of Pediatric Allergology, Dr Von Hauner Children's Hospital, LMU München, Munich, Germany; [5]Comprehensive Pneumology Center Munich (CPC-M), Munich, Germany; [6]German Center for Lung Research (DZL), Munich, Germany; [7]Institute for Medical Information Processing, Biometry and Epidemiology, LMU München, Munich, Germany; [8]Institute of Computational Biology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany; [9]Center for Computational Mathematics, Flatiron Institute, New York, USA;
stefanie.peschel@stat.uni-muenchen.de

Permutation procedures are common practice in statistical hypothesis testing when distributional assumptions about the considered test statistic are not met or unknown. With a small number of permutations, p-values may be either zero [1] or too large to remain significant after adjustment for multiple testing. However, in certain settings, achieving a sufficient number of permutations to obtain accurate p-values is often not feasible. For example, in biomedical studies, the high dimensionality of the data or the use of complex statistical inference methods can make even a single test computationally expensive. A popular heuristic solution to this problem is to approximate extreme p-values by fitting a Generalized Pareto Distribution (GPD) to the tail of the distribution of the permutation test statistics [2]. In practice, however, an estimated negative shape parameter in the GPD combined with extreme observed test statistics can again lead to zero p-values, making subsequent multiple testing problematic.

Here, we propose a complete workflow for accurate and reliable p-value approximation in permutation testing and multiple testing correction. Our framework includes a new method that fits a constrained GPD that strictly avoids zero p-values. We also address the well-known problem of defining an optimal tail threshold for GPD fitting [3] and propose new threshold selection approaches using goodness-of-fit tests. In a multiple testing setting, adjusting the approximated p-values for multiplicity is an essential final step. For this purpose, we introduce a resampling-based False Discovery Rate (FDR) correction procedure that uses the estimated permutation p-values instead of the usual test statistics.

We conduct an extensive simulation study based on the two-sample t-test that demonstrates that our proposed p-value approximation workflow has considerably higher accuracy compared to existing methods. We also illustrate the real-world relevance of our framework in the context of host-associated gut microbiome data analysis, including differential abundance and differential association testing.

Our computational p-value approximation framework, including precise fitting of GPD parameters, tail threshold detection, and multiple testing adjustment, will be made available in the open-source R package permAprox on GitHub and CRAN.

References:

[1] Phipson, Belinda, and Gordon K. Smyth. "Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn." Statistical applications in genetics and molecular biology 9.1 (2010).

[2] Knijnenburg, Theo A., et al. "Fewer permutations, more accurate P-values." Bioinformatics 25.12 (2009): i161-i168.

[3] Langousis, Andreas, et al. "Threshold detection for the generalized Pareto distribution: Review of representative methods and application to the NOAA NCDC daily rainfall database." Water Resources Research 52.4 (2016): 2659-2681.

Wednesday, 06/Sept/2023 9:10am - 9:30am
ID: 462 / **S46: 3**
Presentation Submissions - Featured Session
Featured Sessions: Best practices for Data Monitoring Committees and how to get there

Visualisation and reporting of safety issues

Rachel Phillips
Imperial College London, United Kingdom; r.phillips@imperial.ac.uk

Well-designed graphics are a powerful way to communicate information to an array of different audiences. In randomised controlled trials (RCT) where there is an abundance of complex data on harm they can be a highly effective means to summarise harm profiles and identify potential adverse reactions offering an alternative perspective to the traditional frequency tables. Advances in computer software have improved trialists' capability to produce visualisations; however, little guidance exists on what and how to visually display complex harm data. Therefore with an abundance of visualisation options available and the increasing ease with which they can be implemented we have previously identified researchers' recommendations for visualising harm outcomes with an aim to improve reporting practice in journal articles (1). In this presentation we will extend these ideas to the Data Monitoring Committee (DMC) report. We will present a variety of visualisations dependent on outcome type e.g., binary, count, time-to-event or continuous and the scenario e.g., summarising multiple emerging events or one event of interest. Contrasting these with typical presentation formats (e.g. tables and listings) in DMC reports. We present a decision tree to aid trialists in their choice of visualisations alongside each of the endorsed visualisations, with example interpretation and potential limitations.

1. Phillips R, Cro S, Wheeler G, Bond S, Morris TP, Creanor S, et al. Visualising harms in publications of randomised controlled trials: consensus and recommendations. BMJ. 2022;377:e068983.

Tuesday, 05/Sept/2023 11:20am - 11:40am
ID: 161 / S28: 2
Presentation Submissions - Regular Session (Default)
Topic of Submission: Meta-Analysis and Systematic Reviews
Keywords: Evidence-splitting, Inconsistency, mixed model, arm-based model

A REML method for the evidence-splitting model in network meta-analysis

Hans-Peter Piepho[1], **Johannes Forkman**[2], **Waqas Malik**[1]
[1]University of Hohenheim, Germany; [2]Swedish University of Agricultural Sciences, Uppsala, Sweden; piepho@uni-hohenheim.de

Checking for possible inconsistency between direct and indirect evidence is an important component of network meta-analysis. Recently, an evidence-splitting model has been proposed, that allows separating direct and indirect evidence in a network and hence assessing inconsistency. A salient feature of this model is that the variance for heterogeneity appears in both the mean and the variance structure. Thus, full maximum likelihood (ML) has been proposed for estimating the parameters of this model. ML is known to yield biased variance component estimates in linear mixed models. The purpose of the present paper, therefore, is to propose a method based on residual maximum likelihood (REML). Our simulation shows that this new method is quite competitive to methods based on full ML in terms of bias and mean squared error. In addition, some limitations of the evidence-splitting model are discussed. While this model splits direct and indirect evidence, it is not a plausible model for the cause of inconsistency.

Thursday, 07/Sept/2023 10:40am - 11:00am
ID: 153 / **S67: 1**
Presentation Submissions - Featured Session
Featured Sessions: Industry meets academia: Session in memory of Dieter Hauschke

How to assess bioequivalence of two drugs?

Iris Pigeot
Leibniz Institute for Prevention Research and Epidemiology - BIPS, Germany; pigeot@leibniz-bips.de

How to assess bioequivalence of two drugs?

Dieter Hauschke and I started our collaboration on equivalence trials with the joint supervision of a diploma thesis which resulted in our first joint publication on non-inferiority trials together with Joachim Röhmel and Juliane Schäfer (Pigeot et al. 2003). Following this publication, we worked on four further publications (Hauschke & Pigeot 2005; Röhmel et al. 2005; Hauschke, Steinijans & Pigeot 2007; Pigeot, Hauschke & Shao 2011) – one of those was our book titled *Bioequivalence Studies in Drug Development – Methods and Applications* together with Volker Steinijans which was published by Wiley. Working with Dieter was a wonderful collaboration with a lot of enthusiasm, hard work and … great fun.

This talk is based on the joint book with Dieter. It will therefore discuss various types of equivalence trials with a focus on bioequivalence trials which are of interest when comparing the therapeutic performance of two medicinal products containing the same active substance. Here, we assume that in the same individual similar plasma concentration time courses will result in similar concentrations at the site of action and thus in similar effects, pharmacokinetic data instead of therapeutic results are typically used to demonstrate bioequivalence as an established surrogate marker for therapeutic equivalence. We will distinguish the concepts of average, population and individual bioequivalence and briefly introduce appropriate statistical tests to show bioequivalence of two drugs containing the same active substance.

References

1. Hauschke D, Pigeot I (2005) Establishing efficacy of a new experimental treatment in the "Gold Standard" design. *Biometrical Journal* **47**:782-786 (incl. discussion and rejoinder, p. 797-798)
2. Hauschke D, Steinijans V, Pigeot I (2007) *Bioequivalence Studies in Drug Development – Methods and Applications*. John Wiley & Sons, Ltd, Chichester
3. Pigeot I, Schäfer J, Röhmel J, Hauschke D (2003) Assessing non-inferiority of a new treatment in a three-arm clinical trial including a placebo. *Statistics in Medicine* **22**:883-899
4. Pigeot I, Hauschke D, Shao J (2011) The bootstrap in bioequivalence studies. Invited Paper. *Journal of Biopharmaceutical Statistics* **21**:1126-1139
5. Röhmel J, Hauschke D, Koch A, Pigeot I (2005) Biometrische Verfahren zum Wirksamkeitsnachweis im Zulassungsverfahren. Nicht-Unterlegenheit in klinischen Studien [Biometrical methods for the proof of efficacy in regulatory submissions. Non-inferiority in clinical studies]. *Bundesgesundheitsblatt – Gesundheitsforschung – Gesundheitsschutz* **48**:562-571

Wednesday, 06/Sept/2023 9:50am - 10:10am
ID: 419 / **S44: 4**
Presentation Submissions - Invited Session
Invited Sessions: Online hypothesis testing and subgroup analyses in complex innovative designs
Keywords: adaptive design, interim analysis, multi-stage, blinded sample size recalculation

Multi-stage adaptive enrichment designs with BSSR

Marius Placzek, **Tim Friede**
University Medical Center Göttingen, Germany; marius.placzek@med.uni-goettingen.de

Adaptive enrichment designs offer the possibility to select promising subgroups at (unblinded) interim analyses and reallocate sample size in further stages. We show a design additionally implementing a blinded sample size recalculation (BSSR) in an internal pilot study. The aim is to improve the timing of the interim analysis. To do so we investigate the influence of the timepoint of the sample size review and timepoint of the interim analysis. For normally distributed endpoints, a strategy combining blinded sample size recalculation and adaptive enrichment at an interim analysis is proposed, i.e. at an early timepoint nuisance parameters are reestimated and the sample size is adjusted while subgroup selection and enrichment is performed later. Implications of different scenarios including multiple interim analyses, i.e. multiple stages, and type I error rate control are discussed.

Tuesday, 05/Sept/2023 3:00pm - 3:20pm
ID: 192 / **S33: 3**
Presentation Submissions - Invited Session
Invited Sessions: Endpoints in clinical trials and medical product development: Multiple endpoints, composite endpoints, and biomarkers and surrogate endpoints
Keywords: Nonproportional hazards, Multiplicity

Beyond Proportional Hazards: Multi-Parameter Approaches and Confirmatory Multiple Testing to Quantify Treatment Effects for Time-to-Event Data

Martin Posch
Medical University of Vienna, Austria; martin.posch@meduniwien.ac.at

In randomized controlled clinical trials where time-to-event outcomes are the primary endpoint, hazard ratios are commonly employed to quantify the treatment effect. However, when faced with non-proportional hazards, the hazard ratio is not well defined, necessitating alternative approaches to quantify treatment effects. Such non-proportional hazards might arise, for example, from treatments with delayed effects or when the treatment effect varies across subgroups. In scenarios with non-proportional hazards, a single parameter might be inadequate to capture the differences in survival functions. For instance, in cases where the survival functions intersect, several characteristics of the survival distribution must be considered to determine the more desirable survival function. In this study, we evaluate trials with multiple primary endpoints corresponding to various characteristics of the survival functions. This encompasses the differences and ratios of milestone survival probabilities, differences in the quantiles of the survival distribution, differences in restricted mean survival times, and the average hazard ratio. By employing the counting process representation of survival functions, we show that the parameter estimates are asymptotically multivariate normal and derive their correlations. To account for multiple comparisons, we introduce multiple testing procedures and simultaneous confidence intervals that consider the correlation between different test statistics and also incorporate the logrank test. Through simulations, we assess the finite sample type I error rate and power of the proposed methods and describe the R package 'nph' implementing the procedure.

Reference:

R Ristl, H Götte, A Schüler, M Posch, F König. Simultaneous inference procedures for the comparison of multiple characteristics of two survival functions, 2023 (submitted)

ID: 481 / **STRATOS 2: 1**
Presentation Submissions - Featured Session
Featured Sessions: Ten years of the STRengthening Analytical Thinking for Observational Studies (STRATOS) initiative – progress and looking to the future
Keywords: High-dimensional data, Machine learning, Sample size, Simulation

Statistical analysis of high-dimensional biomedical data: A gentle introduction to analytical goals, common approaches and challenges

Jörg Rahnenführer[1,5], Federico Ambrogi[2], Riccardo De Bin[3], Lisa McShane[4]
[1]Department of Statistics, TU Dortmund University, Dortmund, Germany; [2]Department of Clinical Sciences and Community Health, University of Milan, Milan, Italy; [3]Department of Mathematics, University of Oslo, Oslo, Norway; [4]Biometric Research Program, Division of Cancer Treatment and Diagnosis, National Cancer Institute, Bethesda, MD, USA; [5]for TG9;
rahnenfuehrer@statistik.tu-dortmund.de

Introduction and objectives

The goal of the High-dimensional Data (HDD) Topic Group of the STRATOS initiative (TG9) is to provide guidance amid the jungle of opportunities and pitfalls inherent in the analysis of high-dimensional biological and medical data. Methods for analysis of HDD are rapidly changing, and researchers across different fields, including biostatistics, bioinformatics, and bioengineering, contribute to their development. Advances in statistical methodology and machine learning methods have contributed to improved approaches for data mining, statistical inference, and prediction in HDD settings; however, adoption of these methods has sometimes gotten ahead of understanding of their proper application.

Methods and results

The mission of TG9 includes identification of fundamental principles for analysis of HDD, explanation of available methods, and development of broadly accessible guidance on best practices in this complex and changing landscape. In this talk we present the first published work of TG9 [1]: a comprehensive review aiming to provide a solid statistical foundation for researchers, including statisticians and non-statisticians, who are new to research with HDD or simply want to better evaluate and understand the results of HDD analyses. Following that, we will describe new research topics currently being pursued by TG9. Specifically, guidance materials specific to HDD for sample size calculation, influence and choice of tuning parameters in machine learning applications, and use of plasmode data for simulations are under development. Simulation studies are especially challenging for HDD yet they are essential tools needed to perform evaluation and comparison of different methods.

Conclusions

Proliferation of high-dimensional data in biomedical research has brought unprecedented opportunities to advance knowledge. In order to harness the power of the rapidly evolving repertoire of analysis methods to reveal useful insights from HDD, it is imperative that researchers have access to guidance on the methods available and their proper application.

References

[1] Rahnenführer, J., De Bin, R., Benner, A. et al. Statistical analysis of high-dimensional biomedical data: a gentle introduction to analytical goals, common approaches and challenges. BMC Med 21, 182 (2023). https://doi.org/10.1186/s12916-023-02858-y

Monday, 04/Sept/2023 11:40am - 12:00pm
ID: 380 / **S3: 3**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Machine Learning and Data Science, High dimensional data, genetic and x-omics data
Keywords: Machine learning, Polygenic score, Prediction model

Comparison of classic polygenic scores with machine learning algorithms to predict blood pressure

Tanja K. Rausch, Silke Szymczak, Inke R. König
Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck, Germany; t.rausch@uni-luebeck.de

Blood pressure is a frequently measured clinical parameter, with hypertension being the leading risk factor for the development of cardiovascular disease. Based on the polygenic heritability shown for complex traits like blood pressure, polygenic scores (PGS) are increasingly being used in preclinical and clinical research to stratify individuals according to their genetic susceptibility for targeted prevention, therapy, or prognosis. However, classical PGS use a simple sum of individual genotypes, weighted by the association estimated from single variant genome-wide association studies. Thus, multivariable and non-linear effects are not taken into account. Alternatively, machine learning algorithms can be used for such a score construction.

Machine learning algorithms have not yet been applied to construct polygenic scores to predict blood pressure. Therefore, it is unclear whether more complex algorithms are better able to predict blood pressure than classical scores. This study aims to compare this by using different machine learning algorithms suitable for regression problems such as random forest, linear regression, support vector regression, and k-nearest neighbors regression. For the benchmarking, data from the UK Biobank was used, which is a biomedical database containing genetic and health information from half a million participants from the United Kingdom. The data set was split into a training and a test data set. The training data set was used to generate a simple weighted PGS for blood pressure by performing a genome-wide association study. Moreover, it was used to train different more complex machine learning algorithms. Hyperparameter tuning was performed as well as variable selection where applicable. Prediction performance of the resulting models were compared on the independent test data set by the mean squared error (MSE) and the coefficient of determination (R^2).

The study results provide better insight into whether compressed genetic information obtained by complex machine learning algorithms perform better than classical PGS to predict blood pressure.

Monday, 04/Sept/2023 2:40pm - 3:00pm
ID: 319 / **S10: 3**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Personalized health care
Keywords: Antibiotics resistance, predictive modeling, electronic health records

Developing a predictive model with causal considerations for the risk for antibiotics resistance based on patient health records

Anat Reiner-Benaim
Ben-Gurion University of the Negev, Israel; reinera@bgu.ac.il

In an effort to improve the rational use of antibiotics, many factors have bearing on the rate of resistant pathogens in patients with infection. Taking these factors into account when selecting empirical antibiotic treatment could allow for personalized decisions to be taken, which would result in less unwarranted use of broad spectrum antibiotics. Furthermore, when a patient is presented with a bacterial infection of an unknown pathogen, the physician must decide on prescribing antibiotics before laboratory results are available, thereby imposing uncertainty on the decision.

In this study, we first formulate the decision problem as a causal inference problem and identify the causal effect to be estimated. We then use electronic medical records of over 80,000 hospitalized patients with bacterial infections to develop predictive models for pathogen resistance, and apply causal inference to estimate the effect of antibiotics on future isolation of resistant pathogens.

Monday, 04/Sept/2023 5:30pm - 5:50pm
ID: 400 / **S19: 5**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Free Contributions
Keywords: Early phase trials, Dose-finding, Reporting quality, Guideline, CONSORT

Practical advice on the reporting of statistical items in the new CONSORT extension for early phase dose-finding trials (CONSORT-DEFINE)

Jan Rekowski, Christina Yap
The Institute of Cancer Research, London, United Kingdom; jan.rekowski@icr.ac.uk

CONsolidated Standards Of Reporting Trials (CONSORT) 2010 provides guidance on reporting completed parallel group randomised trials. Using the Enhancing the QUAlity and Transparency Of health Research (EQUATOR) methodological framework for guideline development, CONSORT 2010 has recently been extended to address the special features of early phase dose-finding trials [1]. Such trials have been found to be poorly reported with a detrimental impact on their informativeness and the possibility to make evidence-informed decisions [2]. The resulting consensus-driven CONSORT-DEFINE (CONSORT – DosE-FIndiNg Extension) statement recommends essential items to be included in completed early phase dose-finding trial reports to promote greater transparency, utility of results, and reproducibility. These highly adaptive trials are usually employed as phase I or seamless phase I/II trials and, using dose escalation/de-escalation strategies, they aim to recommend a dosing regimen or a range of dosing regimens for subsequent later phase trials based on safety and other information such as pharmacokinetics, pharmacodynamics, biomarker activity, and clinical activity. They may study any intervention that can be given in different dosages (doses and/or schedules), e.g., drugs, vaccines, cell therapies, gene therapies, digital therapeutics, rehabilitation, or radiotherapy, and may either involve healthy volunteers or people with a condition of interest.

In this presentation, we will emphasize the importance of a widespread implementation of this CONSORT extension, and we will focus on the statistical aspects of new items in CONSORT-DEFINE and items modified from CONSORT 2010 that cover trial design, statistical methods, and analysis. Such items comprise, among others, reporting details on underlying statistical methods for dose escalation/de-escalation strategies and decision-making criteria as well as reporting key outcomes by dosing regimen. We will highlight the importance of including specific items, discuss good examples, and provide practical advice on clear reporting. We hope that CONSORT-DEFINE will ultimately improve participant safety and benefits in early phase dose-finding trials and contribute to transparent reporting while sparing research resources.

Acknowledgement: CONSORT-DEFINE Group and CONSORT-DEFINE Example Guidance Working Group.

[1] Yap, C., et al., The need for reporting guidelines for early phase dose-finding trials: Dose-Finding CONSORT Extension. *Nature Medicine*, 2022. 28(1): p. 6-7.

[2] Yap, C., et al., Assessing the reporting quality of early phase dose-finding trials. *Annals of Oncology*, 2022. 33: p. S24-S24.

Monday, 04/Sept/2023 4:10pm - 4:30pm
ID: 252 / **S20: 1**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Free Contributions
Keywords: Subsampling, Subdata, Polynomial Regression, Design, D-optimality

Optimal Subsampling Design for Polynomial Regression

Torsten Reuter
Otto von Guericke University Magdeburg, Germany; torsten.reuter@ovgu.de

Data Reduction is a major challenge as technological progress has led to a massive increase in data collection to the point where traditional statistical methods fail or computing power cannot keep up. Subsampling reduces data size by selecting a subset from the original data. We study D-optimal subsampling designs for polynomial regression, where the goal is to select a given percentage of the full data that maximizes the determinant of the information matrix. We derive D-optimal subsampling designs under several standard distributional assumptions on the covariates, in particular focusing on the resulting shapes of the subsampling designs. For example, for quadratic regression the D-optimal subsampling design typically has a support of three disjoint intervals. We take a look at the percentage of mass of the optimal subsampling design on the outer intervals compared to the inner one, which changes drastically given the distribution of the covariate, particularly for heavy-tailed distributions. In addition, we examine the efficiency of uniform random subsampling to illustrate the advantage of the optimal subsampling designs. The thus obtained subsampling designs provide simple rules on whether to accept or to reject a data point and therefore allow for an easy algorithmic implementation. We propose a generalization of the Information-Based Optimal Subdata Selection method (IBOSS) to quadratic regression which does not require prior knowledge of the distribution of the covariate and which performs remarkably well compared to the optimal subsampling design. We present an extensive simulation study showing the advantages of our methods over the IBOSS method among others and discuss their computing times. Further we discuss how results extend to other optimality criteria like A- and E-optimality from the Kiefer's Φq-class of optimality criteria, IMSE-optimality for predicting the mean response, or optimality criteria based on subsets or linear functionals of parameters.

Wednesday, 06/Sept/2023 12:00pm - 12:20pm
ID: 181 / **S53: 5**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Real world data and evidence, Time-to-Event Analysis
Keywords: Censored data, Excess Hazard, Net Survival, Relative Survival, Spatial frailty models

Extended Excess Hazard Models for Spatially Dependent Survival Data

André Victor Ribeiro Amaral[1], Francisco Javier Rubio[2], Manuela Quaresma[3], Francisco J. Rodríguez-Cortés[4], Paula Moraga[1]
[1]King Abdullah University of Science and Technology, Saudi Arabia; [2]University College London; [3]London School of Hygiene & Tropical Medicine; [4]Universidad Nacional de Colombia; andre.ribeiroamaral@kaust.edu.sa

Relative survival represents the preferred framework for the analysis of population cancer survival data. The aim is to model the survival probability associated to cancer in the absence of information about the cause of death. Recent data linkage developments have allowed for incorporating the place of residence or the place where patients receive treatment into the population cancer data bases; however, modeling this spatial information has received little attention in the relative survival setting. We propose a flexible parametric class of spatial excess hazard models (along with inference tools), named "Relative Survival Spatial General Hazard" (RS-SGH), that allows for the inclusion of fixed and spatial effects in both time-level and hazard-level components. We illustrate the performance of the proposed model using an extensive simulation study, and provide guidelines about the interplay of sample size, censoring, and model misspecification. Also, we present two case studies, using real data from colon cancer patients in England, aiming at answering epidemiological questions that require the use of a spatial model. These case studies illustrate how a spatial model can be used to identify geographical areas with low cancer survival, as well as how to summarize such a model through marginal survival quantities and spatial effects.

Wednesday, 06/Sept/2023 11:40am - 12:00pm
ID: 409 / **S54: 4**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Software Engineering, Statistical hypothesis testing (covariate adjustment, nonparametric methods, multiple comparisons, …)
Keywords: Sample size calculation, power calculation, online calculator, hypothesis test, exact test

Online sample size calculator

Robin Ristl
Medical University of Viennna, Austria; robin.ristl@meduniwien.ac.at

Deciding on the required sample size for a study is an integral part of study design. The necessary sample size and power calculations are typically performed using dedicated software. Besides commercial software packages, several free to use programs exist, with varying scope of supported hypothesis tests and varying options to accommodate to particular aspects of a planned study such as unequal sample sizes.

A new online sample size calculator was developed with the aim to fit the needs of, both, applied researchers and statisticians involved in study design. The calculator was coded in Java Script and HTML, using distribution functions from the library jStat and adding additional functions for the non-central F-distribution and the non-central hypergeometric distribution. The software is available at https://homepage.univie.ac.at/robin.ristl/samplesize.php and is free to use.

The calculator's interface was designed with the philosophy that only necessary input options should be encountered at first sight and that users should be allowed to enable extended input options when needed.

Regarding continuous outcomes, the calculator currently features the two-sample t-test, paired t-test, analysis of variance with arbitrary number of groups and the Wilcoxon-Mann-Whitney rank sum test. For binary outcomes, tests comparing two proportions are supported, including asymptotic tests and Fisher's exact test, as well as one-sample proportion tests and McNemar's test for paired proportions. Further, calculations for the logrank test for comparing two survival functions and a correlation test based on Fisher's z-transformation are available.

For all tests comparing parallel groups, the calculator allows to calculate sample sizes for unequal allocation ratios between groups, as well as power for unequally sized groups. The calculator has built-in options to consider a Bonferroni-adjustment and to increase the final sample size to match an assumed drop-out rate.

To validate the calculator, results for selected scenarios, covering all function of the software, where compared to results from established commercial software and to independent simulation results.

A current development step is a general implementation of sample size calculation for exact tests. Exact tests typically have a non-monotone power function, i.e. the power may be reduced when the sample size is increased. Thus, an according sample size calculation will have more than one solution and challenges involve the identification of all solutions and a concise presentation to the user. Major sample size tools typically do not include the option to calculate the sample size for exact tests, rather the user is limited to calculate the power for a given sample size.

The aim of the talk is to present the rationale and challenges in developing a sample size tool, to present the validation study and to make accessible the calculator to a wider audience. Further, the particular challenge of sample size calculation for exact tests will be discussed.

Wednesday, 06/Sept/2023 8:30am - 9:10am
ID: 204 / **S44: 1**
Presentation Submissions - Invited Session
Invited Sessions: Online hypothesis testing and subgroup analyses in complex innovative designs
Keywords: multiple testing, online hypothesis testing, platform trial, type I error rate

Online error rate control for platform trials

David Robertson[1], **James Wason**[2], **Franz König**[3], **Martin Posch**[3], **Thomas Jaki**[1,4]
[1]MRC Biostatistics Unit, University of Cambridge, United Kingdom; [2]Newcastle University, United Kingdom; [3]Medical University of Vienna, Austria; [4]University of Regensburg, Germany; david.robertson@mrc-bsu.cam.ac.uk

Platform trials evaluate multiple experimental treatments under a single master protocol, where new treatment arms are added to the trial over time. Given the multiple treatment comparisons, there is the potential for inflation of the overall type I error rate, which is complicated by the fact that the hypotheses are tested at different times and are not necessarily pre-specified. Online error rate control methodology provides a possible solution to the problem of multiplicity for platform trials where a relatively large number of hypotheses are expected to be tested over time. In the online multiple hypothesis testing framework, hypotheses are tested one-by-one over time, where at each time-step an analyst decides whether to reject the current null hypothesis without knowledge of future tests but based solely on past decisions. Methodology has recently been developed for online control of the false discovery rate as well as the familywise error rate. In this talk, we describe how to apply online error rate control to the platform trial setting, present extensive simulation results, and give some recommendations for the use of this new methodology in practice. We also illustrate how online error rate control would have impacted a currently ongoing platform trial.

Thursday, 07/Sept/2023 11:20am - 11:40am
ID: 155 / **S67: 3**
Presentation Submissions - Featured Session
Featured Sessions: Industry meets academia: Session in memory of Dieter Hauschke
Keywords: gold-standard-design, non-inferiority

On intelligent use of the 3-am gold standard design with test treatment, placebo and active control

Joachim Roehmel
Uni Bremen, Germany; joachim.roehmel@gmx.de

Hauschke and Pigeot (2005) initiated a discussion on reasons to use the 3-arm “gold standard” design with experimental treatment, placebo and active control. While these authors focused mainly on the (in their view) dominating role of the active control and developed intelligent statistical analysis strategies, the following discussion from regulators, industry and academia offered a surprising colourful bunch of opinions and worthwhile arguments for the use of the gold standard design, leading to different priorities for the overall judgement of this design as well as on various statistical analysis strategies. Naturally it was possible to incorporate results from the flourishing field of multiple test problems into the analysis. The limits of such research are still to be marked (e.g. B S & S (2022)).

References

1. Establishing Efficacy of a New Experimental Treatment in the ‘Gold Standard’ Design. Dieter Hauschke; Iris Pigeot. Biometrical Journal 47 (2005) , 782–786 and the discussion following this article in the same issue.
2. A Comparison of Multiple Testing Procedures for the Gold Standard Non-Inferiority Trial. Röhmel, J; Pigeot, I: Journal of Biopharmaceutical Statistics, 20: 911–926, 2010
3. Allgemeine Lösungen multipler Testprobleme. Sonnemann E. EDV in Medizin und Biologie 13,120-128, 1982
4. Single-stage, three-arm, adaptive test strategies for non-inferiority trials with an unstable reference Brannath, Scharpenberg, Schmidt. *Statistics in Medicine*. 2022;41:5033–5045

Wednesday, 06/Sept/2023 11:40am - 12:00pm
ID: 189 / **S65: 3**
Presentation Submissions - Invited Session
Invited Sessions: Advancing clinical trial design in rare diseases
Keywords: Small populations, Clinical trial design, Hybrid designs, basket trials

Advancing clinical trial design for small populations: balance and/or rigor?

Kit Roes
Radboudumc, Netherlands, The; kit.roes@radboudumc.nl

The EMA Guideline on Clinical Trials in Small Populations dates back to 2006 and has been “on the list” for revision for at least five years now. The general message that the guideline sets is: On the one hand what could increase efficiency of (randomised) clinical trials in small populations would be equally applicable to large populations. On the other hand, it acknowledges that advancing new therapies is more challenging for small population and thus less common or less tried approaches can be considered, if usual standards are difficult or impossible to implement. From a methodological perspective, several new promising designs have been proposed to improve clinical development – as illustrated in this session. In this presentation I aim to explore how we could address more systematically the potential added uncertainty, trade-offs and also benefits versus the well-known paths with randomised clinical trials. Of the novel designs, I will draw on examples for hybrid designs (leveraging available data) and basket trials. I aim to enhance the discussion on regulatory assessment as well as health economic assessment, to ensure novel designs can be instrumental to new treatments reaching patients with rare diseases.

Wednesday, 06/Sept/2023 8:50am - 9:10am
ID: 425 / **S46: 2**
Presentation Submissions - Featured Session
Featured Sessions: Best practices for Data Monitoring Committees and how to get there
Keywords: iDMC, regulatory, efficay data, safey data, alpha-spending

Regulatory perspective on iDMCs by an experienced iDMC member

Kit Roes
Radboudumc, Netherlands, The; kit.roes@radboudumc.nl

Both the FDA and the EMA have guidance in place on independent Data Monitoring Committees (iDMCs) since 2006, much of which is still relevant and applicable today. More recent developments and potential controversies related to communication have triggered an additional EMA Q&A (1). Additionally, potential functions, roles and specific data needs of iDMCs in adaptive trials and more complex innovative design, such as platform trials, suggest rethinking some of the practices. Against this background, the presentation will address:

- The perspectives on importance and role of iDMC through all stages of drug development.
- Communication in the triangle of sponsor, iDMC and health authorities (including regulatory).
- The essential function of access to comprehensive data (safety and efficacy).
- Considerations of type 1 error, alpha spending and futility in case iDMC have regular access to efficacy data.

The points addressed will take into account relevant experience as iDMC member/chair when impactful advises were to be provided.

Reference:

https://www.ema.europa.eu/en/documents/scientific-guideline/questions-answers-data-monitoring-committees-issues_en.pdf

Wednesday, 06/Sept/2023 11:40am - 12:00pm
ID: 366 / **S53: 4**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Real world data and evidence
Keywords: Real world data, registry, lines of therapy, algorithm, oncology

Rule-based estimation of lines of therapy (LoT) from oncological registry data: the SAKK 80/19 AlpineTIR registry

Alfonso Rojas Mora[1], Caroline Stepniewski[1], Ulf Petrausch[2], Stefanie Hayoz[1]
[1]Competence Center of Swiss Group for Clinical Cancer Research (SAKK), Bern, Switzerland; [2]Onkozentrum Zürich, Zürich, Switzerland; alfonso.rojas@sakk.ch

In oncology, determining the number of prior lines of therapy (LoT) is critical to optimize future treatments, assessing the eligibility for clinical trials and estimate treatment costs. Further, the definition of LoT can have a large effect on the calculation of time-to-event endpoints from clinical data. Due to the complexity of solid cancer treatments, it is challenging to formulate a general framework LoT sequences.

In the SAKK 80/19 AlpineTIR oncological immunotherapy registry, the definition of LoT was extremely heterogeneous across the participating sites, leading to inconsistent and clinically not plausible LoT assignments. Hence, we developed a rule-based approach to determine the LoT from 702 patients enrolled to the AlpineTIR oncological registry, who had received more than 3500 therapies throughout their oncological medical history.

A first approach was to group cancer therapies that were administered within the same period in a LoT, and each newly administered group of treatments would create a new LoT. These simple rules were further refined, to consider, for example, treatment interruptions and drugs that are administered sequentially as part of a single LoT. We calculated the accuracy of the algorithm based on a subset of 352 patients and more than 870 LoT, for which the AlpineTIR coordinating investigator defined the LoT. Additionally, in a subset of 225 patients we compared the lines that were previously defined by the sites to those from the algorithm.

Our simple algorithm has more than 80% accuracy in predicting the LoT that were defined by the coordinating investigator, and the more complex algorithm yielded an accuracy over 90%. To our knowledge, this is the first time that LoT are predicted with such accuracy, providing a good framework to determine LoT from larger and more complex data.

Thursday, 07/Sept/2023 11:00am - 11:20am
ID: 311 / S70: 2
Presentation Submissions - Regular Session (Default)
Topic of Submission: COVID pandemic
Keywords: all-cause mortality, COVID-19, demography, life expectancy, statistical indicator

On the Impacts of the COVID-19 Pandemic on Mortality: Lost Years or Lost Days?

Valentin Rousson, Isabella Locatelli
Center for Primary Care and Public Health (Unisanté), University of Lausanne, Switzerland; valentin.rousson@unisante.ch

Estimating the impact of the COVID-19 pandemic on mortality has been a topic of considerable interest to many scientists and politicians. To quantify this impact, some authors have added up the remaining life expectancies of people who have died with a diagnosis of COVID-19, reaching for example a total of 20.5 million years worldwide in 2020. Although useful for comparing the burden of different diseases, this quantity is however difficult to interpret at face value, due to the lack of a denominator and because it cannot be compared to zero and it is not obvious how to obtain a sensible reference value. In fact, a remaining life expectancy is necessarily greater than zero, even at an advanced age. Another potential issue is that it is based on diagnoses that might be unreliable. This is why many authors have finally attempted to quantify the mortality burden of COVID-19 by simply comparing the (period) life expectancy calculated during the pandemic (e.g. in 2020) with its pre-pandemic level (e.g. in 2019) based on all-cause mortality data (official statistics). This would correspond to the average numbers of years that a hypothetical cohort of people would lose if they lived their entire life under the mortality conditions of 2020 (i.e. with COVID-19) rather than 2019. Given that COVID-19 is expected to soon disappear (or at least become much less virulent), this indicator probably greatly exaggerates its real impact on mortality.

In this presentation, we propose a novel statistical indicator, called “population life loss”, which informs on the average life lost by real (not hypothetical) populations of people living in 2020. This indicator is based on all-cause mortality and demographic data, and can take on positive or negative values, so zero will be a natural reference value here. We calculated population life loss in 2020 for women and men living in 27 countries with available data from the Human Mortality Database. While we could confirm the significant impact of COVID-19 on mortality in 2020 in most countries, it turned out that the estimated population life losses could be counted in days rather than years. For example, while life expectancy loss in 2020 in the United States was of 2.1 years for men and 1.6 years for women, population life loss amounted to 10.1 and 6.7 days, respectively. This should be a useful piece of information from a public health perspective, e.g. to contribute to the delicate debate on the appropriateness of the various restrictive measures taken by governments to fight the pandemic.

Reference: Rousson V, Locatelli I (2022). On the impact of the COVID-19 pandemic on mortality: Lost years or lost days? Frontiers in Public Health 10: 1015501.

Tuesday, 05/Sept/2023 3:20pm - 3:40pm
ID: 404 / **S34: 5**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Meta-Analysis and Systematic Reviews

Investigating the heterogeneity between "study twins"

Christian Röver, **Tim Friede**
University Medical Center Göttingen, Germany; christian.roever@med.uni-goettingen.de

Meta-analyses are commonly performed based on random-effects models, while in certain cases one might also argue in favour of a common-effect (or fixed-effect) model. One such case may be given by the example of two "study twins" that are performed according to a common protocol (or at least very similar study protocols) (see Bender et al, 2018) and which could be considered replications of the same experiment. Here we investigate the particular case of meta-analysis of a pair of randomized controlled trials, focusing on the question of to what extent homogeneity or heterogeneity may actually be discernible based on the data, and including an empirical investigation of published ("twin") pairs of studies. One the one hand, heterogeneity is hard to establish based only on a pair of studies. On the other hand, an empirical sample of "study twins" in fact appears very homogenous (while selection effects might also play a role). Recommendations for meta-analyses of "study twins" will be provided.

Reference:

R. Bender, T. Friede, A. Koch, O. Kuss, P. Schlattmann, G. Schwarzer, G. Skipka. Methods for evidence synthesis in the case of very few studies. Research Synthesis Methods, 9(3):382–392, 2018.

Wednesday, 06/Sept/2023 8:30am - 8:50am
ID: 255 / S47: 1
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Statistical software engineering in the pharmaceutical industry: Increasing productivity, transparency, and reproducibility via open source collaboration
Keywords: open source, software engineering, R packages, collaboration

First year of the Software Engineering working group - working together across organizations

Daniel Sabanes Bove[1], Ya Wang[2], on behalf of the Software Engineering Working Group
[1]Hoffmann-La Roche Ltd, Switzerland; [2]Gilead Sciences Inc., U.S.; daniel.sabanes_bove@roche.com

The Software Engineering (SWE) Working Group (WG) was formed in August 2022 in the American Statistical Association (ASA) Biopharmaceutical Section (BIOP). The SWE WG facilitates cross-organizational collaboration with regular meetings, and currently includes more than 30 members from over 20 organizations.

The primary goal of the SWE WG is to engineer R packages that implement important statistical methods to fill in gaps in the open-source statistical software landscape. The first R package “mmrm” is setting a new standard for fitting mixed models for repeated measures (MMRM) in R.

The secondary goal is to develop and disseminate best practices for engineering high-quality open-source statistical software. The video series “Statistical Software Engineering 101” is introducing specific best practices in accessible format. Furthermore the workshop “Good Software Engineering Practice for R Packages” has been successfully taught in person at a BBS seminar, and the materials are available publicly to train statisticians on best practices.

Communication is key, and the SWE WG was introduced in a BIOP report and maintains a website including a blog at https://rconsortium.github.io/asa-biop-swe-wg.

The SWE WG plans to develop additional new R packages, covering critical and innovative methodology topics in the health-technology assessment (HTA) space, covariate adjustment and Bayesian inference for MMRMs.

We describe the journey of the SWE WG so far and in particular the ingredients for working together successfully, including mutual interest, getting to know each other, and creating mutual trust.

Tuesday, 05/Sept/2023 2:20pm - 2:40pm
ID: 282 / **S29: 2**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Generalized pairwise comparisons
Keywords: Communication, Visual aids, Data visualization

Visuals for Generalized Pairwise Comparisons: innovative tools to explore treatment effects on multiple prioritized outcomes

Samuel Salvaggio, Mickaël De Backer, Vaiva Deltuvaite-Thomas, Sarah Kosta, Emilie Barré, Jean-christophe Chiem, Everardo Saad, Marc Buyse
IDDI, Belgium; samuel.salvaggio@iddi.com

The method of Generalized Pairwise Comparisons (GPC) extends the Wilcoxon Mann-Whitney non-parametric statistical test from a single outcome to multiple outcomes hierarchically ordered for the comparison of two treatment groups (e.g., in a randomized clinical trial). The method estimates a benefit-risk metric called the “Net Treatment Benefit” (NTB), defined as the net probability of a better outcome in one treatment group than in the other. However, properly conveying GPC results is challenging because the method has been recently proposed, making its results unfamiliar and its interpretation not straightforward for clinical-trial stakeholders, including statisticians, physicians, and patients. Additional to its novelty, the multivariate nature of a GPC analysis, while considered a major strength both from a statistical and a clinical point of view, is also a source of challenges for communication around results. This presentation will share several novel ways of communicating GPC results and the NTB to different audiences, from intuitive visual aids that can be quickly understood by non-statisticians to more rigorous and exhaustive tables and figures.

Wednesday, 06/Sept/2023 11:20am - 11:40am
ID: 394 / **S54: 3**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Software Engineering
Keywords: Julia, Design of Experiments, Bayesian Statistics

A Julia Package for Bayesian Optimal Design of Experiments

Ludger Sandig
Technische Universität Dortmund, Germany; sandig@statistik.tu-dortmund.de

Suppose a toxicologist wants to study the effects of a drug. But at which concentration(s) and at which point(s) in time should they take measurements? Mathematically, this question can be framed as a problem of optimal experimental design. For the underlying nonlinear regression model this means selecting covariate values and corresponding samples sizes such that the observations are as informative as possible about the unknown model parameters. One way to formalize this is the D-criterion: Maximize the expected gain in Shannon information with respect to some prior density, or equivalently minimize the average volume of a confidence region around the maximum likelihood estimate of the model parameters. Because the optimal number of design points is not known beforehand, designs are represented as probability measures. The resulting optimization problem is computationally intensive, and it is typically solved using particle-based global optimization heuristics.

Existing free and open-source software packages for this task have several drawbacks. Typically, only a small number of response functions and design criteria are implemented. Further ones can be added only by the maintainer of the package, and fiddly interfacing with external C/C++ code is necessary. Where user-defined functions can be supplied, they run much slower than pre-packaged ones. This is especially notable when the covariate has more than one dimension, e.g. in dose-time-response models. Moreover, package code is often not well modularised, making it hard for third parties to contribute extensions. For these reasons it is difficult to adapt existing software for complex experimental setups.

In this talk, we present a Julia package that addresses these issues. Julia is a scientific high-level dynamic programming language with a performance comparable to statically compiled C. Julia's type system and multiple dispatch mechanism allow for a concise implementation that elegantly reflects the optimization problem's mathematical structure. We demonstrate its flexibility on a wide range of examples.

Thursday, 07/Sept/2023 9:30am - 9:50am
ID: 200 / **S57: 3**
Presentation Submissions - Invited Session
Invited Sessions: Causal inference and the art of asking meaningful questions

Optimal regimes for algorithm-assisted human decision-making

Aaron Leor Sarvet
EPFL, Switzerland; aaron.sarvet@epfl.ch

Foundational work on causal inference and dynamic treatment regimes presents a promising pathway towards precision medicine. In a precision-medicine system, decision rules might be algorithmically individualized based on an optimal rule previously learned from non-experimental or experimental data. However, there is some resistance to the notion that implementation of an optimal regime, successfully learned from the data, will result in better expected outcomes on average, compared to existing human-decision rules: care providers may be inclined to override the treatment recommendations provided by the identified optimal regimes, based on their privileged patient observations. In this talk, I will review existing methodology for learning optimal regimes and clarify the validity of the care provider's skepticism. Then, I will present methodology for leveraging human intuition by identifying a *super*-optimal regime using data generated by either nonexperimental or experimental studies, and clarify when a fusion of such data is beneficial. The superoptimal regime will indicate to a care provider -- in an algorithm-assisted decision setting -- precisely when expected outcomes would be maximized if the care provider would override the optimal regime recommendation and, importantly, when the optimal regime recommendation should be followed regardless of the care-provider's assessment.

Tuesday, 05/Sept/2023 5:30pm - 5:50pm
ID: 166 / **S38: 4**
Presentation Submissions - Invited Session
Invited Sessions: From multivariate to high-dimensional and functional data
Keywords: Multivariate Data, Correlation matrices, Nonparametric Testing, Bootstrap

Testing Hypotheses about Correlation Matrices in General MANOVA Designs

Paavo Sattler, Markus Pauly
TU Dortmund, Germany; paavo.sattler@tu-dortmund.de

Covariance and correlation matrices are essential tools for investigating random vectors' dispersion and dependency structure. Especially
correlation matrices that are not affected by units allow investigating the dependency structures of random vectors or comparing them.
We introduce an approach for testing various null hypotheses that can be formulated based on the correlation matrix. Examples cover MANOVA-type hypothesis of equal correlation matrices as well as testing for special correlation structures such as, e.g., sphericity. Apart from existing fourth moments, our approach requires no other assumptions, allowing applications in various settings. To improve the small sample performance, a bootstrap technique is proposed and theoretically justified, as well as some Taylor-expansion-based approaches.

ID: 482 / **STRATOS 1: 1**
Presentation Submissions - Featured Session
Featured Sessions: Ten years of the STRengthening Analytical Thinking for Observational Studies (STRATOS) initiative – progress and looking to the future
Keywords: Observational studies, guidance, STRATOS initiative

Experience and progress with developing guidance for the analysis of key topics in observational research

Willi Sauerbrei[1,6], Michal Abrahamowicz[2], Saskia Le Cessie[3], Marianne Huebner[4], Ruth Keogh[5], James Carpenter[5]
[1]Medical Center - University of Freiburg, Germany; [2]Department of Epidemiology and Biostatistics, McGill University, Montreal, Canada; [3]Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands; [4]Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824, USA; [5]Department of Medical Statistics, London School of Hygiene & Tropical Medicine, London, UK; [6]for the STRATOS initiative; wilhelm.sauerbrei@uniklinik-freiburg.de, james.carpenter@lshtm.ac.uk

The STRATOS initiative was launched at ISCB 2013 and the first STRATOS paper summarized the motivation, mission, structure and aims of this international initiative (Sauerbrei, Abrahamowicz, Altman, le Cessie and Carpenter, 2014, www.http://stratos-initiative.org/). Providing accessible, evidence-based guidance for key topics in the design and analysis of observational studies is the main aim. Guidance is intended for applied statisticians and other data analysts with varying levels of statistical background and experience. The focus is on health sciences research, but the content is also relevant for applications of statistics in other empirical sciences.

In 2013 STRATOS started off with seven topic groups (TGs) focusing on different aspects of study design and analysis methodology (1- Missing data, 2- Selection of variables and functional forms in multivariable analysis, 3- Initial data analysis, 4- Measurement error and misclassification, 5- Study design, 6- Evaluating diagnostic tests and prediction models, 7- Causal inference). For their specific topic, each group provided a brief summary of the state of research, main issues, main aims and planned future research (Sauerbrei et al., 2014). Two further TGs were initiated in 2015 on the topics of Survival analysis (TG8) and High-dimensional data (TG9). Summaries are available on the STRATOS website.

To coordinate the activities of the initiative, and to help improve standards of both methodological and applied research, we started several cross-cutting panels, that work on issues common to all TG's, including simulation, visualization, and most recently about open science.

In this talk we will provide a short introduction illustrating the necessity of guidance for analysis of observational studies and outline experience and progress of the STRATOS initiative.

Thursday, 07/Sept/2023 9:10am - 9:30am
ID: 346 / **S60: 3**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …)
Keywords: Continuous variable, fractional polynomial, Influential point, model building, sample size, simulated data

Effects of Influential Points and Sample Size on the Selection and Replicability of Multivariable Fractional Polynomial Models

Willi Sauerbrei, Edwin Kipruto
Medical Center - University of Freiburg, Germany; wilhelm.sauerbrei@uniklinik-freiburg.de

Background: The multivariable fractional polynomial (MFP) approach combines variable selection using backward elimination with a function selection procedure (FSP) for fractional polynomial (FP) functions. It is a relatively simple approach which can be easily understood without advanced training in statistical modelling. For continuous variables, a closed test procedure is used to decide between no effect, linear, FP1 or FP2 functions. Influential points (IPs) and small sample sizes can both have a strong impact on a selected function and MFP model.

Methods: We used simulated data with six continuous and four categorical predictors to illustrate approaches which can help to identify IPs with an influence on function selection and the MFP model. Approaches use leave-one or two-out and two related techniques for a multivariable assessment. In eight subsamples we also investigated the effects of sample size and model replicability, the latter by using three non-overlapping subsamples with the same sample size. For better illustration, a structured profile was used to provide an overview of all analyses conducted.

Results: The results showed that one or more IPs can drive the functions and models selected. In addition, with small sample size, MFP was not able to detect some non-linear functions and the selected model differed substantially from the true underlying model. However, when the sample size was relatively large and regression diagnostics were carefully conducted, MFP selected functions or models that were similar to the underlying true model.

Conclusions: For smaller sample size, IPs and low power are important reasons that the MFP approach may not be able to identify underlying functional relationships for continuous variables and selected models might differ substantially from the true model. However, for larger sample sizes a carefully conducted MFP analysis is often a suitable way to select a multivariable regression model which includes continuous variables. In such a case, MFP can be the preferred approach to derive a multivariable descriptive model.

Tuesday, 05/Sept/2023 4:50pm - 5:10pm
ID: 201 / S41: 3
Presentation Submissions - Regular Session (Default)
Topic of Submission: Clinical trial designs (adaptive and platform designs, external controls, …), Real world data and evidence
Keywords: real world data, hybrid control, propensity score weighting

Augmenting randomized trials with real-world data: a simulation study evaluating methods for hybrid control arm analyses

Rafael Sauter[1], Benjamin Ackerman[1], Martina Fontana[1], Ignazio Craparo[2], Brian Hennessy[1]
[1]Johnson & Johnson; [2]Alira Health, Italy; rsauter1@its.jnj.com

Randomized controlled trials (RCTs) are considered the gold standard for estimating causal effects of new therapies and interventions, yet statistical challenges remain in detecting treatment effects among rare disease populations, particularly when serious outcomes also occur infrequently. In such cases, innovative approaches exist to supplement trials with evidence from real-world data (RWD) by constructing a hybrid control arm, retaining the benefits of randomization while increasing study sample size and power. Identifying suitable RWD for this use case is critical, and data appropriateness are highly dependent on the design of the candidate RCT, its study population, and its primary outcome measure. Even with high-quality RWD, differences in study populations may exist, and must be properly accounted for to ensure comparability. In this work, we present propensity score-type weighting methods to align study populations when conducting hybrid control arm analyses.

When augmenting a RCT control arm with RWD in a hybrid control analysis, it is important to first identify any baseline patient covariate imbalances. When prognostic factors of the outcome are imbalanced between the two studies, then naively pooling the control arms together without any population adjustment could result in a biased treatment effect estimate. Such imbalances can be accounted for with propensity scores by modeling the probability of study membership conditional of the observed prognostic factors and weighting the RWD patients by the odds of their propensity score. This ensures that the distribution of covariates in the weighted RWD sample is more similar to the demographic profile of the RCT. The outcome model is then fit on the combined data of RCT and weighted RWD patients. When all prognostic factors are accounted for, and the propensity score model is correctly specified, the augmented treatment effect is unbiased.

In our proposed analysis method, the hybrid control group consists of both unit-weighted RCT patients and propensity score-weighted RWD patients. The scale of the RWD weights is a function of the study sample sizes, and if one study is much larger than the other, inflated variability among the combined control arm could result in lower power and type-1 error. To address this, we assess methods to rescale the RWD weights.

Operating characteristics of the proposed methods are established in a simulation study, where RCT and RWD samples are generated with baseline covariates that are increasingly prognostic of the outcome and are also increasingly imbalanced between the studies. In doing so, we illustrate conditions where use of the proposed methods to augment trials with RWD yield unbiased estimates with greater precision than RCT-only analyses while maintaining the type-1 error. We provide guidance on how to adequately quantify covariate imbalance and how to accordingly justify the appropriateness of a propensity score approach. We highlight key criteria when selecting suitable RWD based on study population comparability and highlight practical considerations and limitations when implementing the proposed methods for hybrid control arm analyses.

Tuesday, 05/Sept/2023 11:20am - 11:40am
ID: 138 / **S26: 2**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Clinical trial designs (adaptive and platform designs, external controls, …), Statistical hypothesis testing (covariate adjustment, nonparametric methods, multiple comparisons, …)

Simultaneous confidence intervals for an extended Koch-Röhmel design in three-arm non-inferiority trials

Martin Scharpenberg, **Werner Brannath**
University of Bremen, Germany; mscharpenberg@uni-bremen.de

Three-arm `gold-standard' non-inferiority trials are recommended for indications where only unstable reference treatments are available and the use of a placebo group can be justified ethically. For such trials several study designs have been suggested that use the placebo group for testing 'assay sensitivity', i.e. the ability of the trial to replicate efficacy. Should the reference fail in the given trial, then non-inferiority could also be shown with an ineffective experimental treatment and hence becomes useless. In this talk we extend the so called Koch-Röhmel design where a proof of efficacy for the experimental treatment is required in order to qualify the non-inferiority test. While efficacy of the experimental treatment is an indication for assay sensitivity, it does not guarantee that the reference is sufficient efficient to let the non-inferiority claim be meaningful. It has therefore been suggested to adaptively test non-inferiority only if the reference demonstrates superiority to placebo and otherwise to test δ-superiority of the experimental treatment over placebo, where δ is chosen in such a way that it provides proof of non-inferiority with regard to the reference's historical effect. We extend the previous work by complementing its adaptive test with compatible simultaneous confidence intervals.

Confidence intervals are commonly used and suggested by regulatory guidelines for non-inferiority trials. We show how to adopt different approaches to simultaneous confidence intervals from the literature to the setting of three-arm non-inferiority trials and compare these methods in a simulation study. Finally we apply these methods to a real clinical trial example.

Tuesday, 05/Sept/2023 2:20pm - 2:40pm
ID: 165 / S30: 2
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), Time-to-Event Analysis
Keywords: Gradient Boosting, Interactions, Model trees, Pseudo-values, Survival probabilities

Pseudo-Value Regression Trees

Alina Schenk, Moritz Berger, Matthias Schmid
Institute of Medical Biometry, Informatics and Epidemiology, Medical Faculty, University of Bonn; schenk@imbie.uni-bonn.de

This work presents a semi-parametric modeling technique for estimating the survival function from a set of left-truncated and right-censored time-to-event data. Our method, named pseudo-value regression trees (PRT), is based on the pseudo-value regression framework, modeling individual-specific survival probabilities by computing pseudo-values and relating them to a set of covariates. The standard approach to pseudo-value regression is to fit a main-effects model using generalized estimating equations (GEE). PRT extend this approach by building a multivariate regression tree with pseudo-value outcome and by successively fitting a set of regularized additive models to the data in the nodes of the tree. Due to the combination of tree learning and additive modeling, PRT are able to perform variable selection and to identify relevant interactions between the covariates, thereby addressing several limitations of the standard GEE approach. In addition, PRT includes time-dependent covariate effects in the node-wise models. Interpretability of the PRT fits is ensured by controlling the tree depth. Based on the results of two simulation studies, we investigate the properties of the PRT method and compare it to several alternative modeling techniques. Furthermore, we illustrate PRT by analyzing survival in 3,652 patients enrolled for a randomized study on primary invasive breast cancer.

Thursday, 07/Sept/2023 8:30am - 8:50am
ID: 157 / **S62: 1**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Epidemic short-term forecasting in real time
Keywords: COVID-19, ensemble forecast, forecast evaluation, real-time forecasting, Forecast Hub

Collaborative forecasting of COVID-19 in Germany and Poland

Melanie Schienle[1,2], **Daniel Wolffram**[1,2], **Johannes Bracher**[1,2]
[1]Karlsruhe Institute of Technology, Germany; [2]Heidelberg Institute for Theoretical Studies; melanie.schienle@kit.edu

Short-term forecating of infectious diseases can contribute to situational awareness and resource planning during infectious disease outbreaks. During the COVID-19 pandemic, such forecasts have received considerable attention, and it is increasingly acknowledged that multi-model systems can improve the reliability of results. So-called Forecast Hubs have therefore been launched in various countries in order to coordinate modelling efforts, enable a coherent comparative evaluation of different models and their combination into a forecast ensemble. In the opening talk of this session, we will provide an overview of the German and Polish COVID-19 Forecast Hub, which we operated from May 2020 through April 2021, when it was merged into a larger effort led by the European Centers for Disease Prevention and Control. We will discuss practical aspects of coordinating real-time forecasting with numerous independent research teams as well as statistical and epidemiological challenges we encountered. Particular attention will be given to a pre-registered evaluation study which we conducted between October 2020 and April 2021. The results indicate that while deaths can be predicted with some success, case forecasts are very challenging, and in particular abrupt changes in trends are difficult to predict. Ensemble forecasts overall showed good relative performance, for most forecast targets outperforming most or all individual models.

Tuesday, 05/Sept/2023 3:20pm - 3:40pm
ID: 405 / S35: 5
Presentation Submissions - Regular Session (Default)
Topic of Submission: Epidemiology, High dimensional data, genetic and x-omics data
Keywords: Genetics, Transcriptomics, Proteomics, GWAS, conditional causal inference

Integrated transcriptome- and proteome-wide association studies nominate causal determinants of kidney function

Pascal Schlosser[1], Jingning Zhang[1], Hongbo Liu[2], Aditya L. Surapaneni[1], Eugene P. Rhee[3], Dan E. Arking[4], Bing Yu[5], Eric Boerwinkle[5,6], Paul Welling[4], Nilanjan Chatterjee[1], Katalin Susztak[2], Josef Coresh[1], Morgan E. Grams[1,7]

[1]Johns Hopkins Bloomberg School of Public Health, USA; [2]Perelman School of Medicine, University of Pennsylvania, USA; [3]Massachusetts General Hospital, USA; [4]Johns Hopkins University School of Medicine, USA; [5]University of Texas Health Science Center at Houston, USA; [6]Baylor College of Medicine, USA; [7]New York University Grossman School of Medicine, USA; pschlos3@jhu.edu

Background: The pathophysiological causes of chronic kidney disease development are not fully understood. Genome-wide association studies (GWAS) can be utilized as a foundation for causal inference to identify molecular pathways involved in the pathogenesis. One of the key challenges is the translation of genetic variants associated with disease to the respective physiological relevant units – the genes. While transcriptome-wide association studies rooted in genetic instruments have advanced our capabilities for causal inference using gene expression (e.g. PrediXcan and FUSION approaches) the joint analysis of related transcriptomic and proteomic features is an unmet statistical challenge.

Results: We applied a Mendelian Randomization technique that focused on instrumental genetic variables within the respective gene region (*cis*-variants) and allows us to conduct a genome-wide screen in a two-sample design. We combined tissue-specific weights derived by elastic net regression in the GTEx project for transcriptome-wide association studies (TWAS) in relevant tissues (kidney cortex, kidney tubule, liver, and whole blood), plasma-specific weights derived for proteome-wide association studies (PWAS), and the most recent genome-wide association study (GWAS) summary statistics for three markers of kidney filtration (glomerular filtration rate (GFR) estimated with serum creatinine, GFR estimated by serum cystatin C, and blood urea nitrogen) and one of kidney damage (albuminuria). This allowed us to assess the effects of 12,893 genes and 1,342 proteins on each of the kidney markers. We found 1,561 significant associations (Bonferroni adjusted) distributed among 260 genomic regions that were supported by TWAS and/or PWAS as putatively causal. We then prioritized 153 of these genomic regions using additional Bayesian colocalization analyses (posterior probability >80%). These findings intertwined with local co-regulation of neighboring genes left us with an unmet statistical challenge of the joint analysis of the different tissues and neighboring genes. To integrate the genetic influence on the transcriptome and the proteome, we used regression with a summary statistics approach to incorporate the *cis*-regulated genetic correlation of the different models, performing conditional causal inference to identify the underlying tissue of an association as well as to evaluate whether multiple independent signals were contained within the same genomic region. Our findings were supported by existing knowledge (e.g., animal models for *MANBA*, *DACH1*, *SH3YL1, INHBB*), exceeded the underlying GWAS signals (28 region-trait combinations without significant GWAS hit), were confirmed by experimental follow up in a clinical cohort (INHBC kidney disease progression hazard ratio=1.86, CI=1.37-2.52) and differentiated markers of kidney filtration from those with roles in creatinine and cystatin C metabolism.

Conclusion: In summary, this study combined multimodal, genome-wide association studies to generate a catalog of putatively causal target genes and proteins relevant to disease. We extended the causal inference approach to allow for conditional analysis to prioritize tissues, and demonstrated the application based on the example of kidney function and damage which can guide follow-up studies in physiology, basic science, and clinical medicine.

Wednesday, 06/Sept/2023 8:50am - 9:10am
ID: 354 / **S49: 2**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Time-to-Event Analysis
Keywords: treatment switching, multistate model, Cox regression

Long term safety evaluations in the presence of switching: evaluation of two approaches

Sandra Schmeller[1], Rima Izem[2], Byron Jones[2], Valentine Jehl[2]
[1]Institute of Statistics, Ulm University, Germany; [2]Novartis Pharma AG, Basel, Switzerland; sandra.schmeller@uni-ulm.de

Further investigation of the long-term benefit and risk profile of a new therapy after approval in a real-world setting is essential to ensure relevant information is made available to patients and health care providers for the safe use of a drug. Our goal is to improve the understanding of two analysis methods of post-marketing safety studies (1). The motivating case study is a post-marketing observational study for a biologic drug investigating potential long-term risk of malignancy. The main methodological challenge is to answer a comparative safety question, in the real-world setting, where subjects switch away or to the investigational drug during their follow-up. Standard approaches include an “ever/never exposed” type of design where the interpretation becomes challenging in the presence of frequent treatment switching. Further declinations of this approach have been proposed, in which time and events assignment to respective treatment arms differ. In the so called “experimental hierarchical approach”, the experimental arm is given a higher ranking and time and events occurring after initiation of the investigating drug are assigned to the investigating drug irrespective of further treatment change. In the “overlapping approach”, on the contrary, both arms are handled equally and given same importance. Both approaches attempt to evaluate a treatment effect accounting for possible treatment switch. We explain these two estimators with the help of multistate model methodology and investigate both approaches under the Nullhypothesis of no group difference. This leads to a discussion of the timescale and the rational is that the estimators are biased under the Nullhypothesis. Finally, we show a possible improvement to be considered. A simulation study will evaluate the type 1 error rate and power of the different analytical methods under different switching and incidence scenarios.

Reference:

Joel M Kremer, Clifton O Bingham III, Laura C Cappelli, Jeffrey D Greenberg, Ann M Madsen, Jamie Geier, Jose L Rivas, Alina M Onofrei, Christine J Barr, Dimitrios A Pappas, et al. Postapproval comparative safety study of tofacitinib and biological disease-modifying antirheumatic drugs: 5-year results from a united states–based rheumatoid arthritis registry. ACR open rheumatology, 3(3):173–184, 2021.

Wednesday, 06/Sept/2023 12:00pm - 12:20pm
ID: 229 / **S55: 5**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Machine Learning and Data Science, Software Engineering
Keywords: Manufacturing, Scheduling, Nonclinical, Simulation

Improving Production Capacity and Asset Utilization of Biologics Drug Product Lines Through Simulation

Christian Schmid, Ilmari Ahonen
F. Hoffmann-La Roche AG, Switzerland; christian.schmid.cs8@roche.com

Scheduling a production plan for a drug product line comes with many challenges. The planner has to account for the available equipment, working times of operators, and scheduled maintenance operations. Other constraints are related to the products being manufactured, as each product has a specific format and process duration. In addition, there can also be unplanned events which require a total redesign of the schedule.

In order to facilitate the planning process, we have developed a data-driven simulation tool that accounts for all of these constraints and provides explicit planning schedules that can be easily geared to the needs of each product line. The simulation can identify ideal production plans for different priorities, e.g., schedules that maximize the overall equipment effectiveness (OEE), minimize the end-to-end lead time, or even decrease the amount of manual labor between the change over of different products. In addition, this simulation can use historical data to evaluate the chances of successfully running a given production plan and, in consequence, identify potential bottlenecks. It can also help process managers to quickly adapt and optimally assign tasks especially during unplanned events.

We have successfully scaled and applied our method to multiple lines at Roche. The results will be shared and lessons learned will be discussed. The simulation tool helps increase the asset utilization and production capacity, which ultimately enables more drug products to be delivered to our patients.

Thursday, 07/Sept/2023 12:00pm - 12:20pm
ID: 156 / S67: 5
Presentation Submissions - Featured Session
Featured Sessions: Industry meets academia: Session in memory of Dieter Hauschke

When safety data meet survival analysis

Claudia Schmoor, Martin Schumacher
Faculty of Medicine and Medical Center, University of Freiburg, Germany; claudia.schmoor@uniklinik-freiburg.de

One of Dieter Hauschkes main interests was the benefit assessment of medical interventions. From 2010 onwards, he yearly organized GMDS Workshops on this topic which led to the re-establishment of the GMDS Arbeitsgruppe Therapeutische Forschung (ATF). His aim was to bring together colleagues from academia, industry, regulatory and HTA institutions for discussing their different perspectives. During GMDS 2014 in Göttingen, the specific workshop topic was “Statistical methods for the analysis of adverse event data” with talks published in a special issue of Pharmaceutical Statistics (2016).

Analysis of safety data in terms of adverse events (AE) is an essential part of the evaluation of therapies in clinical trials and, especially, in their benefit assessment. Traditionally, statistical analyses based on simple tables presenting incidence proportions have been the standard approach. However, such analyses do not take into account time-related issues including frequently occurring problems such as varying follow-up times, censoring, and competing events. Although methods derived from survival analysis were already suggested many years ago, such approaches have only rarely been applied.

As head of ATF, Dieter was one of the initiators of the joint project group of the ATF (GMDS) and the Arbeitsgruppe Pharamazeutische Forschung (APF) of the IBS German Region on the topic „Analyse unerwünschter Ereignisse bei variablen Beobachtungszeiten in der Nutzenbewertung“ in 2016. As a first step, this group published recommendations for AE analyses (Unkel et al, 2019). The second step was an interdisciplinary joint venture between academia and industry called SAVVY (Survival analysis of Adverse eVents with VarYing follow-up times) that will be the main focus of our contribution to this session (Stegherr et al, 2021a, 2021b, 2021c, Rufibach et al, 2022).

The SAVVY project aims to improve the analysis of AE in clinical trials through the use of survival analysis techniques appropriately dealing with varying follow-up times leading to censoring and competing events. The main purpose is to illustrate the amount of empirical bias of estimators typically used to quantify AE risk, as incidence proportion, probability transform of incidence density, and the Kaplan-Meier estimator in comparison to the Aalen-Johansen estimator as the gold-standard for estimating AE probabilities P(AE). Estimators were compared for the analysis of 17 clinical trials (186 types of investigated AEs) from ten sponsor organizations descriptively and more formally using random effects meta-analysis. It was demonstrated that the resulting bias can be substantial, i.e. considerable underestimation of P(AE) by incidence proportion, and considerable overestimation of P(AE) by Kaplan-Meier. The SAVVY project is ongoing with planned further activities for promoting the correct methods as investigations in specific disease areas, illustrations in medical journals, provision of programming code, and the ultimate aim of inducing updates of ICH and other guidelines dealing with safety analyses.

Reference:

1. Kieser et al (2016), Pharm Stat 15:290–323.
2. Unkel et al (2019), Pharm Stat 18:166-183.
3. Stegherr et al (2021a), Biom J 63:650-670.
4. Stegherr et al (2021b), Pharm Stat 20:1125-1146.
5. Stegherr et al (2021c), Trials 22:420, 2021.
6. Rufibach et al (2022), Stat Biopharm Res online.

Monday, 04/Sept/2023 4:30pm - 4:50pm
ID: 332 / **S20: 2**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Clinical trial designs (adaptive and platform designs, external controls, …), Preclinical drug development, safety and toxicology
Keywords: Optimal Designs, Nonlinear Regression Modelling, Alert Concentrations

Optimal design for identifying alert concentrations

Kirsten Schorning[1], **Kathrin Möllenhoff**[2]
[1]TU Dortmund University, Dortmund, Germany; [2]Heinrich Heine University, Düsseldorf, Germany; schorning@statistik.tu-dortmund.de

The determination of alert concentrations, where a pre-specified threshold of the response variable is exceeded, is an important goal of concentration-response studies. Recently, several model-based testing procedures were developed that provide the identification of alerts at concentrations, which were not measured during the study. These model-based approaches are based on the fits of nonlinear concentration-response curves and therefore their quality strongly depends on the set of concentrations at which observations were taken.

In this talk, we address the optimal design problem for the identification of alert concentrations in order to improve these model-based testing procedures with respect to their power. Consequently, an optimal design minimizes the maximum variance of the estimator of potential alert concentration. Optimal design theory (equivalence theorem, efficiency bounds) is developed for this design problem and the results are illustrated in several examples identifying the alert concentration under the assumption of different dose-response relationships. In particular, it is demonstrated within a simulation study that using the optimal design results in more powerful tests for identifying alerts than using other commonly used “non-optimal” designs.

Monday, 04/Sept/2023 11:00am - 11:20am
ID: 145 / **S5: 1**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Finding the right dose – Project Optimus and beyond
Keywords: Oncology Dose Escalation, Back-fill Cohorts, Bayesian Logistic Regression Model, Project Optimus

Back-fill cohorts in oncology phase I trials: implementation, practical considerations, and impact on operating characteristics

Lukas Schröter
Boehringer Ingelheim Pharma GmbH & Co. KG, Germany; lukas.schroeter@boehringer-ingelheim.com

Traditionally, oncology Phase I dose escalation trials focus on identifying a maximum tolerated dose and on the generation of safety evidence. Here, a Bayesian Logistic Regression Model is a popular and widely used method to guide dose escalation.

We currently experience a huge paradigm shift in oncology related to dose optimization ("Project Optimus"). With the emergence of modern cancer therapies, the rationale of simply moving forward with the maximum tolerated dose up to a registrational trial could lead to suboptimal dose levels and is not accepted anymore. With this, it is of interest to generate more evidence across a broad dose range already during the earliest phases of clinical development to allow for a better characterization of activity and tolerability. To achieve this, the FDA recommends in their draft guidance to consider "*add[ing] more patients to dose-level cohorts in a dose-finding trial […]*". In dose escalation trials, this concept is known as back-fill cohorts, where additional patients are recruited on lower dose levels while dose escalation is ongoing at higher doses. While back-fill cohorts help generating more evidence already during dose finding, this approach comes with various statistical and clinical challenges that need to be addressed.

In this presentation, we will discuss some of these challenges that occur when back-fill cohorts are implemented in dose-finding trials guided by a Bayesian logistic regression model. This includes possible conditions on when such cohorts should be opened and how this additional data can be included in the statistical analyses. Additionally, we will present operating characteristics from trial simulations that investigated the impact of enrolling back-fill cohorts during dose escalation in trials with monotherapy and/or combination therapy arms. Special focus lies on the risk of exposing a high numbers of patients to overly toxic doses.

Monday, 04/Sept/2023 2:00pm - 2:20pm
ID: 397 / **S14: 1**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Clinical trial designs (adaptive and platform designs, external controls, …)
Keywords: non-reproducibility, Phase II, Phase III, selection bias, adjustment methods

Non-reproducibility between phase II and III: Region selection and go/no-go related bias and methods for its correction

Kristin Schultes[1,2,3], Heiko Götte[3]
[1]University of Applied Science Fulda, Germany; [2]Master's degree Program, Medical Biometry/Biostatistics, University of Heidelberg, Germany; [3]Merck Healthcare KGaA, Germany; kristin.schultes@external.merckgroup.com

Non-reproducibility between phase II and III: Region selection and go/no-go related bias and methods for its correction

Kristin Schultes[a,,b,c], Heiko Götte[c]

a University of Applied Science Fulda, Germany

b Master's degree Program, Medical Biometry/Biostatistics, University of Heidelberg, Germany

c Merck Healthcare KGaA, Darmstadt, Germany

Background: High failure rates in Phase III trials are a major challenge in clinical research. Often, promising treatment effects in Phase II cannot be reproduced in the following Phase III study. This lack of reproducibility might be related to the two main differences between Phase II and Phase III trials: Phase II being conducted in a limited number of (selected) regions or sites – opposed to global phase III trials – and Phase III only being conducted after observing promising results in phase II (go/no-go decision rules). Both aspects can be sources of overoptimism/bias. Suitable methods for bias adjustment are needed. Additive and multiplicative adjustment was proposed for go/no-go related bias, and the approximate-Bayesian-computation (ABC) approaches were used for addressing different forms of selection bias. The objective was to define a quantitative approach for bias adjustment due to selection (region and go/no-go) for the Phase II treatment effect estimates and corresponding probability of success (PoS) estimates for phase III.

Method: The go/no-go decision after Phase II is based on observed treatment effects leading to a bias between the observed and true effect. In contrast to that, region selection in Phase II is performed before any patient data is collected. Therefore, to describe characteristics of the different sources of bias and to compare properties of the different effect estimates, an additional type of bias is introduced: the "true effect bias" between the selected regions' true effect and the true overall effect (composed of all potential regions for phase III). In a simulation study, the different adjustment methods are examined according to the level of bias reduction in treatment effects and estimated PoS for three bias scenarios: Only decision rule, only region selection or both.

Results: Each of the adjustment methods correct for overestimation. Adjusted treatment effect estimates are generally preferable to non-adjusted estimates. While the additive and multiplicative adjustment methods are appropriate in specific combinations of the true overall effect, the number of events, the region selection, and/or the decision rule, the ABC-adjusted treatment effect estimates perform best across all combinations. The adjusted PoS shows less consistency over simulation scenarios.

Discussion: Stricter go/no-go rules increase bias but reduce (false) decisions to go to Phase III. In contrast to that, more extreme region selection increases bias and go decisions at the same time. Therefore, in drug development processes, the risk of a region bias should be addressed. ABC methods can adjust for region bias and seem useful for practice.

Monday, 04/Sept/2023 5:30pm - 5:50pm
ID: 436 / **S20: 5**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …)
Keywords: Sum scores, item response theory, psychometrics, latent variable modeling

Using Item response theory for testing assumptions underlying clinical scores

Daniel Schulze, Ulrike Grittner
Charité - Universitätsmedizin Berlin, Germany; daniel.schulze2@charite.de

Scores are frequently used in medicine to aggregate patient features or to sum up questionnnaires. Group comparisons on such scores are just as common. Relatively little attention is paid to the assumptions of simple unweighted aggregations across several characterstics. Item response theory (IRT) is a framework to understand four implicit assumptions that are made in scores: 1) the aggregated features (or items) reflect a single underlying cause, 2) the features contribute equivocally to the score, 3) the features are measured without any error, and 4) the measurement properties are exactly the same in two groups whose comparison is of interest to the researcher. Violations of these assumptions result in potentially severe bias in an effect of interest, e.g. in group mean comparisons. Depending on which assumption is violated, bias can diminish or exaggerate an effect. We thus advocate testing these assumptions by means of IRT modeling. We will discuss the concept of latent variables and IRT in its application to medical research. We introduce the most common models and discuss model parameters, model testing, and measurement invariance. IRT concepts are introduced with the help of real data from the Danish alcohol and drug consumption survey and are accompanied by necessary programming in R. We will discuss pitfalls and limitations.

Thursday, 07/Sept/2023 10:40am - 11:00am
ID: 218 / **S68: 1**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Design of preclinical experiments
Keywords: Optimal design, gene expression data

Designs for the simultaneous inference of concentration-response curves

Leonie Schürmeyer, Kirsten Schorning, Jörg Rahnenführer
Faculty of Statistics, TU Dortmund University, Germany; schuermeyer@statistik.tu-dortmund.de

Understanding the concentration-response relationship of a candidate drug is one of the main goals in toxicology and especially drug development. Therefore, several authors emphasized the importance of optimal designs regarding concentration-response experiments. Using classical optimal design approaches significantly enhances the precision of the estimated concentration-response curve or rather specific parameters. Thus, optimal designs can substantially improve comprehension in toxicological context.
We extend classical approaches of optimal design of experiments (see [Pukelsheim] among many others) such that they can be applied for the analysis of concentration-response relationships in the context of gene expression data. The experimental conditions of such data are new challenges in planning those experiments, since all genes are elevated simultaneously. Thousands of concentration-response relationships need to be determined, so the key question is which design works best for the simultaneous analysis. Thereby gene expression data of valproic acid applied to human embryonic stem cells is analyzed to compare different designs [Krug et al.]. First of all genes with biologic activity are evaluated with the Multiple Comparison Procedure and Modelling approach [Bretz et al.]. Simultaneously the true underlying concentration-response relationships are fitted using separate sigmoid Emax models for all active genes. Then locally D-optimal designs are identified for every considered gene. Based on the locally D-optimal design, a summarized design is established using the K-means algorithm. Moreover, a D-optimal design for simultaneous inference is developed by using an appropriate Bayesian D-optimality criterion. Both developed designs are compared to the design, originally used in the experimental set-up, an equidistant and log-equidistant design with respect to their D-efficiencies. Moreover, a simulation study is conducted to demonstrate the differences of all designs practically. In that simulation study, we also vary the total sample size to investigate its influence on the precision of the model fits.

The results actively demonstrate to support the consideration of using optimal design approaches for gene expression data. Especially the D-optimal design for simultaneous inference and the design developed with K-means perform considerably better than the originally used design and the log-equidistant design, both in terms of the theoretical and the practical comparison. Measured by the root mean squared error (RMSE) the original, equidistant and log-equidistant design have an inferior performance in the simulation study compared to the two other designs. Besides the precision of the model fits highly increases by enlarging the total sample size. Thus, it is recommendable using as many observations as possible, whereupon minimal 27 data points should be used in this analysis. Summarizing, the D-optimal design for simultaneous inference leads to the most exact model fits, consequently, it should be preferred to the other designs investigated.

[Bretz et al.] Bretz, F., Pinheiro, J. C. and Branson, M. (2005): Combining multiple comparisons and modeling techniques in dose-response studies, Biometrics, 61(3), p. 738-748.

[Pukelsheim] Pukelsheim, F. (2006): Optimal Design of Experiments, Wiley, New York.

[Krug et al.] Krug, A. K. et al. (2013): Human embryonic stem cell-derived test systems for developmental neurotoxicity: a transcriptomics approach. Arch Toxicol, 87, p. 123–143.

Thursday, 07/Sept/2023 8:30am - 8:50am
ID: 466 / S59: 1
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Volume-outcome relationships in health care
Keywords: #minimum-volume-thresholds #federal-joint-committee germany

Minimum Volume Thresholds of the Federal Joint Commit-tee in Germany

Horst Schuster
GKV-Spitzenverband, Germany; horst.schuster@gkv-spitzenverband.de

Based on correlation between the volume components and medical outcome minimum volume thresholds for specific medical treatments have been implemented by the Federal Joint Committee (Gemeinsamer Bundesausschuss, G-BA) in Germany. Decision making in-cludes both, the catalogue of procedures and the mode of participation in the delivery of enclosed procedures, respectively. The prospective permission to perform enlisted proce-dures relies on a prognosis about surpassing the defined threshold in the following year to be based on the case number of the previous year. According to this prospective ap-proach new thresholds unfold there effects prior to that. The contribution explains opera-tion mode of the rules enacted by the G-BA and provides first insides on the effects.

Wednesday, 06/Sept/2023 8:30am - 8:50am
ID: 123 / **S48: 1**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Clinical trial designs (adaptive and platform designs, external controls, …), Statistical hypothesis testing (covariate adjustment, nonparametric methods, multiple comparisons, …)
Keywords: sample size re-estimation, nonparametric tests, relative effect, conditional power

Sample Size Re-estimation for the Wilcoxon-Mann-Whitney and Brunner-Munzel Test

Stephen Schüürhuis[1], Tobias Mütze[2], Georg Zimmermann[3], Frank Konietschke[1]
[1]Institute of Biometry and Clinical Epidemiology, Charité - Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany; [2]Statistical Methodology, Novartis Pharma AG, Basel, Switzerland; [3]Team Biostatistics and Big Medical Data, IDA Lab Salzburg, Paracelsus Medical University, Salzburg, Austria;
stephen.schueuerhuis@charite.de

Proper sample size determination throughout the planning stage of a randomized controlled trial is crucial. The sample size is usually determined as the (minimum) sample size to detect an alternative δ, say, with target power of $1-\beta$ (e.g. 80%) at significance level α (e.g. 5%). In general, however, the sample size computed does not only depend on the aforementioned parameters, but also on nuisance parameters (e.g. variance). Hence, the appropriateness of the resulting sample size particularly depends on the validity of the specified input assumption on the effect.

In practice, however, a-priori knowledge about parameter values might be scarce, e.g. in novel indications or rare diseases. Accordingly, the assumptions on the effect (and its variability) can be highly uncertain. Therefore, allowing for modifications of the preplanned sample size during the trial based on updated knowledge about this effect might be an attractive alternative option. Interim sample size re-estimation is regarded as a particular class of interim adaptions within the general framework of adaptive designs (see, e.g., [1]). In (unblinded) sample size re-estimation designs, a first-stage cohort of patient data can be used in order to adaptively de- or increase the overall sample size based on the interim effect estimate as compared to the initially planned overall sample size.

While extensive theory has been developed for binary, continuous and survival endpoints (e.g. [1, 2]), there has been comparatively little discussion in the adaptive design literature on how to perform sample size re-estimation if the underlying statistical procedure is nonparametric, e.g. if the analysis should be done using the Wilcoxon-Mann-Whitney or the Brunner-Munzel-test. In disease areas such as amyotrophic lateral sclerosis (ALS), however, rank-based methods are commonly used, see e.g. [3], and they are considered more robust if distributional assumptions or asymptotics do not hold. The effect size of those tests is the so-called relative effect $p = P(X<Y)+0.5P(X=Y)$ for $X\sim F_X$,$Y\sim F_Y$, where $p>0.5$ means that Y stochastically tends to larger values than X.

In this talk, we will present unblinded sample size re-estimation procedures for the Wilcoxon-Mann-Whitney and the Brunner-Munzel test. In particular, we will focus on interim sample size adaptations based on estimates of the relative effect p utilizing the conditional power of those tests, i.e. the probability to obtain a significant result given the already observed interim data. Moreover, we will provide simulation studies to investigate the designs with respect to type I error rate control in various settings for continuous and ordered categorical data.

[1] G. Wassmer and W. Brannath. Group Sequential and Confirmatory Adaptive Designs in Clinical Trials.Springer, Heidelberg, 2016.

[2] C. Chuang-Stein, K. Anderson, P. Gallo, and S. Collins. Sample size reestimation: a review and recommendations. Drug information journal: DIJ/Drug Information Association, 40(4):475–484, 2006.

[3] J. D. Berry, R. Miller, D. H. Moore, M. E. Cudkowicz, L. H. Van Den Berg, D. A. Kerr, Y. Dong, E. W. Ingersoll, and D. Archibald. The combined assessment of function and survival (cafs): a new endpoint for als clinical trials. Amyotrophic lateral sclerosis and frontotemporal degeneration, 14(3):162–168, 2013.

Tuesday, 05/Sept/2023 11:00am - 11:20am
ID: 148 / S28: 1
Presentation Submissions - Regular Session (Default)
Topic of Submission: Meta-Analysis and Systematic Reviews
Keywords: meta-analysis, publication bias, funnel plot, asymmetry, simulation

LFK index does not reliably detect bias in meta-analysis

Guido Schwarzer, Gerta Rücker
Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center - University of Freiburg;
guido.schwarzer@uniklinik-freiburg.de

Background: The *LFK* index has been introduced and promoted as a new improved quantitative method to detect bias in meta-analysis (Furuya-Kanamori et al., 2018). Putative main advantage compared to established tests for funnel plot asymmetry (Sterne et al., 2011) is that its performance does not depend on the number of studies in the meta-analysis (Furuya-Kanamori et al., 2020). An independent evaluation of the *LFK* index has not been conducted to our knowledge.

Methods: We conducted a simulation study under the null hypothesis of no bias in meta-analysis with a continuous normally distributed outcome, comparing the *LFK* index test to three standard tests for funnel plot asymmetry (Sterne et al., 2011). In total, 108 scenarios were evaluated by varying the number of studies, mean and standard deviation in experimental group and between-study heterogeneity. In addition, two settings with smaller or larger group sample sizes were considered.

Results: In general, the type I error of the *LFK* index test showed a massive dependency on the number of studies *k* ranging from 30 percent (*k*=10) to about 2 percent (*k*=100) for smaller group sample sizes and from 19 percent (*k*=10) to 0 percent (*k*=100) for larger group sample sizes. Egger's test adhered only well to the prespecified significance level of 10 percent under homogeneity, but was too liberal (smaller groups) or conservative (larger groups) under heterogeneity. The rank test was too conservative for most simulation scenarios. The Thompson-Sharp test was too conservative under homogeneity, but adhered well to the significance level in case of heterogeneity.

Conclusion: The *LFK* index in its current implementation should not be used to assess bias in meta-analysis. The Thompson-Sharp test shows the best performance in heterogeneous meta-analyses.

References:

1. Furuya-Kanamori, L., Barendregt, J. J. & Doi, S. A. R. A new improved graphical and quantitative method for detecting bias in meta-analysis. *International Journal of Evidence-Based Healthcare* **16**, 195–203 (2018).
2. Furuya-Kanamori, L. et al. P value–driven methods were underpowered to detect publication bias: analysis of Cochrane review meta-analyses. *Journal of Clinical Epidemiology* **118**, 86–92 (2020).
3. Sterne, J. A. C. et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ* **343**, d4002 (2011).

Monday, 04/Sept/2023 4:50pm - 5:10pm
ID: 342 / **S17: 3**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Prognostic and predictive biomarkers in personalized medicine
Keywords: machine learning, feature selection, multiplicity, predictive biomarkers, subgroup identification

Using knockoffs for controlled predictive biomarker identification

Kostas Sechidis[1], **Matthias Kormaksson**[1], **David Ohlssen**[2]
[1]Novartis, Switzerland; [2]Novartis, US; kostas.sechidis@novartis.com

One of the key challenges of personalized medicine is to identify which patients will respond positively to a given treatment. The area of subgroup identification focuses on this challenge, that is, identifying groups of patients that experience desirable characteristics, such as an enhanced treatment effect. A crucial first step towards the subgroup identification is to identify the baseline variables (eg, biomarkers) that influence the treatment effect, which are known as predictive variables. Many subgroup discovery algorithms return importance scores that capture the variables' predictive strength. However, a major limitation of these scores is that they do not answer the core question: “Which variables are actually predictive?” With our work we answer this question by using the knockoff framework, which is a general framework for controlling the false discovery rate when performing *prognostic* variable selection. In contrast, our work is the first that uses knockoffs for *predictive* variable selection. We introduce two novel knockoff filters: one parametric, building on variable importance scores derived from a penalized linear regression model, and one non-parametric, building on causal forest variable importance scores. We conduct extensive simulations to validate performance of the proposed methodology and we also apply the proposed methods to data from a randomized clinical trial. This talk is based on our recent paper: https://onlinelibrary.wiley.com/doi/full/10.1002/sim.9134

Thursday, 07/Sept/2023 11:00am - 11:20am
ID: 110 / S67: 2
Presentation Submissions - Featured Session
Featured Sessions: Industry meets academia: Session in memory of Dieter Hauschke
Keywords: Bioequivalence, Concurrent Control, Randomisation, Target Population

What can bioequivalence studies teach us about clinical trials?

Stephen Senn
Stephen Senn, United Kingdom; stephen@senns.uk

When I was a statistician in the pharmaceutical industry in the late 1980s and early 1990s a particular development in asthma on which I worked was plagued by the necessity of switching formulations. We started with a solution aerosol, switched to a suspension aerosol and then to a dry powder formulation and finally tried to introduce a multi-dose version of the dry powder formulation of a particular beta-agonist on which we were working. Comparative 'bridging' studies for the formulations were necessary and although we could not carry out conventional bioequivalence studies, much of the methodology of such studies was relevant. In reading the literature to increase my understanding of this field, a name I frequently came across was that of Dieter Hauschke. Later, I was fortunate enough to get to know him well and also very pleased that he was prepared to contribute a book, together with Volker Steinijans (who had collaborated with Dieter for many years) and Iris Pigeot to the Statistics in Practice series that I edited.

As a tribute to Dieter I have chosen bioequivalence as a topic. However, I shall not review the extensive controversies on the analysis of such studies, a subject on which Dieter worked extensively but rather issues such studies raise for clinical trials more widely. Two in particular are important. First, whether it is a requirement that clinical trials should be carried out in representative subjects. If so, then this is a requirement which bioequivalence studies spectacularly fail, since they are nearly always carried out in healthy volunteers. Second, what the role of blinding is in clinical trials when the purpose is to assert equivalence.

As regards the first, I shall argue that recent claims that there can be substantial formulation-by-sex effects and that therefore care should be taken to ensure that women are adequately included in such trials are misguided. What is important, however, is to use an appropriate scale for analysis and this applies more widely to therapeutic trials for which we should understand that representativeness is not a pre-requisite for transportability and in any case not adequately addressed by inclusion criteria.

As regards the second, I shall argue that in equivalence studies, blinding is valuable but nonetheless does not provide the protection that it does in studies designed to show superiority.

Both issues are related to concurrent control. I shall further argue that some proposals in the recent literature on causal analysis are ignoring study effects and that these are plausibly important and that their existence undermines a number of claims that have been made regarding the use of observational studies.

Thursday, 07/Sept/2023 11:40am - 12:00pm
ID: 431 / **S64: 4**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Estimands and causal inference, Free Contributions
Keywords: patient-centric, virtual clinical trials, decentralization component, estimand attributes

Decentralized clinical trials: scientific considerations through the lens of the estimand framework

Nikolaos Sfikas
Novartis, Switzerland; nikolaos.sfikas@novartis.com

While the industry and regulators' interest in decentralized methods is long-standing, the Covid-19 pandemic accelerated and broadened the adoption and the experience with these methods in clinical trials. The key idea in decentralization is bringing the clinical trial design, typically on-site, closer to the patient's experience (on-site or off-site). Thus, potential benefits of decentralizing studies include reducing the burden of participation in trials, broadening access of clinical trials to a more diverse population, or using innovative endpoints collected off-site.

This presentation facilitates evaluation of the added value and the implications of decentralized designs beyond the operational aspects of their implementation. The proposed approach is to use the ICH E9(R1) estimand framework to guide the strategic decisions around each decentralization component. Furthermore, the framework can guide clinical trialists to systematically consider the implications of decentralization, in turn, for each attribute of the estimand. Illustration of the use of this approach with a decentralized trial case study will show that the proposed systematic process can uncover the scientific opportunities, assumptions and potential risks associated with a possible use of decentralized components in the design of a trial. This process can also highlight the benefits of specifying estimand attributes in a granular way. The presentation will hence demonstrate that bringing a decentralization component into the design not only impacts estimators and estimation but can also correspond to addressing more granular questions, thereby uncovering new target estimands.

Monday, 04/Sept/2023 4:50pm - 5:10pm
ID: 226 / **S16: 2**
Presentation Submissions - Invited Session
Invited Sessions: Causal discovery with a view to the life sciences

Consistent and efficient mixed integer programming for causal discovery

Ali Shojaie
University of Washington, United States of America; ashojaie@uw.edu

Learning the structure of directed acyclic graphs (DAGs), known as causal discovery, is computationally and statistically challenging. We cast the problem as a mixed-integer program with an objective function composed of a convex quadratic loss function and a regularization penalty subject to linear constraints. The optimal solution to this mathematical program is known to have desirable statistical properties under certain conditions. However, the state-of-the-art optimization solvers are not able to obtain provably optimal solutions to the existing mathematical formulations for medium-size problems within reasonable computational time. To address this difficulty, we tackle the problem from both computational and statistical perspectives. Computationally, we propose an efficient mixed-integer quadratic optimization (MIQO) model, the layered network formulation. In addition to offering improvements compared with the existing approaches, the new formulation can also take advantage of easily obtainable super structures, such as the moral graph, to reduce the number of possible DAGs. Statistically, we propose an early stopping criterion to terminate the branch-and-bound process in order to obtain a near-optimal solution to the mixed-integer program, and establish the consistency of this approximate solution.

Tuesday, 05/Sept/2023 11:20am - 11:40am
ID: 232 / **S27: 2**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Statistical issues in health care provider comparisons
Keywords: Double robustness, Procedural quality indicators, Pseudo-values, Restricted mean survival time

Comparing procedural quality indicators of health care across regions using restricted mean survival time

Hana Šinkovec[1,2], Walter Gall[1], Georg Heinze[1]
[1]Center for Medical Data Science, Medical University of Vienna, Austria; [2]Biotechnical Faculty, University of Ljubljana, Slovenia; hana.sinkovec@gmail.com

Background:

Practice guidelines (GL) synthesize the available scientific evidence and assist health care providers by recommending management strategies for patients with a given condition. Quantifying and comparing the adherence to such recommendations has become increasingly important. A procedural quality indicator (PQI) describes one aspect of a recommended treatment which can be assessed by evaluating the trace of a patient in linked routine data bases typically available to social insurance institutions. For example, patients with myocardial infarction should receive continuous supply of high-power statins after discharge from hospital. A pure binary evaluation of this PQI (patient fully adhered to GL or not) may miss the variation in the extent of partial compliances. Moreover, patients may not be observable for full 12 months after their index event because of reinfarction, death or loss-to-follow-up.

Objective:

In order to overcome the deficiencies of a binary evaluation of PQI, we propose to estimate the expected time in compliance via restricted mean survival time (RMST). We demonstrate its application by comparing patients' time in compliance to continuous high-power statin therapy over 12 months after MI across the political districts of Austria.

Methods:

RMST naturally accommodates right-censoring and competing risks. RMST can be estimated using pseudo-values, an approach that lends itself to formal causal comparisons, e.g., by doubly robust estimation.

Results:

Our registry data set consisted of 26,821 patients, residing in 116 political districts of Austria, who were discharged between January 2012 and December 2014 with a principal diagnosis of acute MI. Analysis used 12 months as the restriction time and was adjusted for patients' demographic characteristics and high-dimensional information based on billed hospital stays and filled drug prescriptions collected from a period of three months before the index MI event. RMST was estimated using pseudo-values, considering the time from index MI event until end of continuous supply with high-power statins, death or reinfarction, or end of the PQI period of 12 months, whatever came first. Death and reinfarction were considered as competing events. Results revealed considerable variation in compliance across political districts of Austria.

Conclusions:

The RMST provides an estimate of patients' expected time in compliance which has a clear interpretation and is easily communicable. In applications when time in PQI is naturally restricted, it can also be expressed as a fraction of the total time covered by the PQI. Unlike alternative methods for time-to-event analyses the RMST can summarize the difference between two regions when non-compliance initially diverges and later converges. Existing methods and available software allow to model the RMST with covariates directly without a proportional hazards or any distributional assumption; these methods also facilitate the causal comparisons in the presence of right-censoring and competing events.

Thursday, 07/Sept/2023 9:10am - 9:30am
ID: 186 / S57: 2
Presentation Submissions - Invited Session
Invited Sessions: Causal inference and the art of asking meaningful questions
Keywords: Bias, causal inference, confounding, E-value, sensitivity analysis

Are E-values too optimistic or too pessimistic? Both and neither!

Arvid Sjölander, Sander Greenland
Karolinska Institutet, Sweden; arvid.sjolander@ki.se

The E-value is a popular tool to assess the sensitivity to unmeasured confounding; the higher the E-value, the stronger the unmeasured confounding must be to "explain away" an observed association. However, despite its popularity, the E-value has also been heavily debated and criticized. Ioannidis et al (2019) argued that a high E-value may give an unwarrented optimistic impression, since the accumulated effect of many unmeasured confounders may be large and "trump" even a high E-value, even though the effect of each separate cofounder is small. In contrast, Greenland (2020) argued that a small E-value may give an unnecessarily pessimistic impression, since bias by unmeasured confounders may be weakened considerably due to their associations with measured confounders. These criticisms may appear contradictory, and leave the reader wondering whether E-values should be interpreted as being too optimistic or too pessimistic. In this presentation we will attempt to reconcile these criticisms, and use a real data example to argue that both interpretations are valid. The presentation will be largely non-technical, and focus on fundamental conceptual issues.

Tuesday, 05/Sept/2023 4:10pm - 4:50pm
ID: 213 / **S38: 1**
Presentation Submissions - Invited Session
Invited Sessions: From multivariate to high-dimensional and functional data

Functional data analysis on the example of analysis of variance

Łukasz Smaga
Adam Mickiewicz University, Poznan, Poland; ls@amu.edu.pl

Functional data analysis (FDA) is a branch of statistics that analyzes observations treated as functions, curves, or surfaces. To represent the data in such a way, one needs only to measure some variable over time or space, which is a scenario encountered in many fields, such as brain imaging data, medical measurements over time, biological development, etc. Then the discrete data observed at so-called design time points can be transformed into functional data. Such a representation allows us to avoid many problems of classical multivariate statistical methods, for example, the curse of dimensionality and missing data. Therefore, numerous methods have been developed for classification, clustering, dimension reduction, regression, and statistical hypothesis testing for functional data. The methods are based on different approaches, for example, dimension reduction (basis expansion, functional principal components), random projections on multivariate data, and aggregating pointwise statistics. During the talk, we present the different aspects and strategies of the functional data analysis methods with a special focus on the functional analysis of variance. The latter covers, in particular, one-way, multi-way, univariate, multivariate, independent observations, and repeated measurements.

Monday, 04/Sept/2023 11:00am - 11:20am
ID: 109 / **S7: 1**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical hypothesis testing (covariate adjustment, nonparametric methods, multiple comparisons, …)
Keywords: multiple, testing, behrens-fisher

Nonparametric methods for clustered data in the several sample case

Erin Sprünken
Charité - Universitätsmedizin, Germany; erin-dirk.spruenken@charite.de

In many trials and experiments, subjects are not only observed once, but multiple times, resulting in a cluster of possibly correlated observations. For example, mice sharing the same cage, patients providing two brain hemispheres or students of the same class are typical examples of clustered data.

Typically, under the assumption of normally distributed data, mixed models are used for analysis. However, this model assumption is rather strict and hard to justify in most real data analyses. Furthermore, skewed data (e.g. waiting times), discrete data (e.g. count data) or ordered categorical data measured on an ordinal scale are typical endpoints in a variety of trials. This motivates the use of nonparametric methods which do not rely on any specific data distribution.

For the two-sample case, several nonparametric procedures exist. For binary clustered data, a chi-square-test for contingency tables can be used. Furthermore, generalizations of the Wilcoxon-Mann-Whitney-test exist for testing the null hypothesis of equal distributions of clustered data. An extension is provided by a procedure under a less strict null hypothesis formulated in terms of the Wilcoxon-Mann-Whitney effect.

Here, we aim to generalize the procedures for the analysis of several samples. Thus, we propose a general nonparametric framework for comparing multiple groups of clustered data under mild assumptions. We present different inference methods, namely ANOVA-type test statistics and a multiple contrast test procedure and investigate their asymptotic behavior. Extensive simulation studies indicate that the methods control the type-1 error rate well, even with small sample sizes. A real data example illustrates the application of the proposed methods.

Monday, 04/Sept/2023 2:40pm - 3:00pm
ID: 257 / **S13: 3**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …)
Keywords: Nonlinearity, interactions, functional forms, tree-based modeling, varying coefficients

A tool to detect nonlinearity and interactions in generalized regression models

Nikolai Spuck, Matthias Schmid, Moritz Berger
Institute for Medical Biometry, Informatics and Epidemiology, Medical Faculty, University of Bonn, Germany; spuck@imbie.uni-bonn.de

Generalized linear models (GLMs) are a popular tool for regression analysis. They are based on the assumption that the relationship between the modeled outcome of interest and the covariates is linear. In addition, it is frequently assumed that the effect of a covariate is independent of the values of other covariates, neglecting possible interactions. These assumptions, however, may be too restrictive in many applications and lead to biased effect estimates. There are numerous alternative approaches for modeling continuous covariates like categorization, polynomial regression, generalized additive models (GAMs) and tree-based methods. However, while the application of variable selection methods in regression analysis has become increasingly common, methods that provide guidance regarding the choice of suitable functional forms for continuous covariates are still lacking. To address this issue, we propose an algorithm that examines various modeling alternatives and is able to detect nonlinearity and interactions between covariates if they are present. The algorithm utilizes tree-based splits which makes the resulting effects easily interpretable. More specifically, it indicates whether (i) linear effects are sufficient (indicating the use of a simple GLM), (ii) varying linear effects should be included in the model formula, (iii) one or several covariates exhibit non-linear effects (calling for the use of a GAM), or (iv) interaction effects occur in the data (hinting that the use of a tree-based method may be beneficial). We illustrated the algorithm by an application to data from patients who suffered from chronic kidney disease. The performance of the algorithm was assessed based on detection rates in a simulation study. Results of the simulation study indicate that the algorithm is able to proficiently detect nonlinearity and identify the correct functional form for a continuous covariate in settings with medium to high sample sizes and moderate noise. Some specific interactions structures were less likely to be identified correctly.

Tuesday, 05/Sept/2023 12:20pm - 12:40pm
ID: 469 / **S25: 5**
Presentation Submissions - Featured Session
Featured Sessions: Young Statisticians Sessions and Panel Discussion
Keywords: applied statistics, proteomics, multiple membership

Statistical methods for the analysis of massspectrometry data with multiple membership

Mateusz Staniak[1], Jurgen Claesen[2], Tomasz Burzykowski[3], Małgorzata Bogdan[1], Olga Vitek[4]
[1]University of Wrocław, Poland; [2]Epidemiology and Data Science, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands; [3]Hasselt University, Belgium; [4]Northeastern University, USA; mateusz.staniak@uwr.edu.pl

Mass spectrometry (MS) is a core technology for proteomics. It allows for the identification and quantification of proteins in biological samples. In a mass spectrometry experiment, peptides – smaller fragments of proteins - are ionized, separated based on their mass and charge, and quantified. Resulting data are complex and high-dimensional: they include up to thousands of proteins and tens of thousands of peptides. This type of data is important in drug discovery and other stages of drug development, and in various fields of proteomics research.

Typically, mass spectrometry data are used to estimate relative abundance of proteins in different biological conditions. One of the major challenges in the analysis of mass spectrometry is the protein inference problem: deriving a list of proteins that are present in the sample based on identified peptides. It is complicated due to three factors: false identifications of peptides based on mass spectra, presence of peptide sequences that can be attributed to multiple proteins (shared peptides) and one-hit wonders - proteins identified only by a single peptide. Similarly, protein quantification - estimating the relative abundance of proteins in different conditions - is difficult in the presence of shared peptides, as it is not clear how to distribute peptide abundance among their respective proteins. From statistical perspective, inclusion of shared peptides in models that are used to estimate protein abundances from peptide-level data introduces the multiple membership structure, in which observations (peptide intensities) may belong to multiple groups defined by proteins.

Typically, shared peptides are removed from analysis of MS data, which leads to loss of peptide-level information and lack of ability to estimate abundances of proteins that are identified only by shared peptides or by a single unique peptide. Our goal is to propose a statistical methodology capable of including shared peptides in downstream MS data analysis to increase the number of proteins that can be identified and quantified reliably, and improve the power of statistical analysis. In this talk, we will present two classes of non-linear models that can be used to describe labeled and label-free mass spectrometry experiments with shared peptides: models with peptide-specific weights and non-weighted models. In labeled experiments, multiple biological conditions or subjects may be measured jointly. In this case, peptides have natural quantitative profiles, which we use to estimate the degree of their protein membership (weights). We use these weights to estimate protein-level summaries of peptide data, which are then used for comparisons of biological conditions. For non-labeled experiments, we will present an approach that uses additional information from raw spectra to enable protein quantification with shared peptides. We will illustrate proposed models with biological data, and provide analytical and simulation study-based results on their statistical properties.

Research presented in this talk was done in collaboration with Genentech company, Northeastern University (USA) and Hasselt University (Belgium), and was financially supported by the National Science Center grant 2020/37/N/ST6/04070 (Poland).

Tuesday, 05/Sept/2023 4:10pm - 4:30pm
ID: 183 / **S37: 1**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Causal estimands for time to event data
Keywords: Estimands, Hazard Ratios, Survival Analysis, Competing Events, Causal Inference

On the choice of estimands in clinical trials with time-to-event outcomes

Mats Stensrud
Ecole Polytechnique Fédérale de Lausanne, Switzerland; mats.stensrud@epfl.ch

In this presentation, I will discuss how to formulate and choose an estimand, beyond the marginal intention to treat effect, from the perspective of a decision maker and drug developer. I will specifically consider randomized clinical trials with time-to-event outcomes. I will emphasize that a careful articulation of a practically useful research question should either reflect decision making at this point in time or future drug development. A common feature of estimands that are practically useful is that they correspond to possibly hypothetical but well-defined interventions in identifiable (sub)populations. To illustrate my points, I will consider examples from clinical trials involving competing, recurrent and intercurrent events.

Wednesday, 06/Sept/2023 12:00pm - 12:20pm
ID: 461 / **S51: 5**
Presentation Submissions - Featured Session
Featured Sessions: IBS-DR/ROeS Award session
Keywords: Optimal treatment regimes, Decision making, Clinical medicine, Unmeasured confounding

Optimal treatment regimes assisted by algorithms

Mats Stensrud
Ecole Polytechnique Fédérale de Lausanne, Switzerland; mats.stensrud@epfl.ch

Doctors and other care providers desire to implement decision rules that, when applied to individuals in the population of interest, yield the best possible outcomes. For example, the current focus on precision medicine reflects the search for individualized treatment decisions, adapted to a patient's characteristics. In this presentation, I will consider how to formulate, choose and estimate effects that guide individualized treatment decisions. In particular, I will introduce a class of regimes that are guaranteed to outperform conventional optimal regimes. I will further argue that identification of these "superoptimal" regimes and their values requires exactly the same assumptions as identification of conventional optimal regimes in several common settings. The superoptimal regimes can also be identified in data fusion contexts, in which experimental data and (possibly confounded) observational data are available. The performance of the superoptimal regimes will be illustrated in two clinical examples.

Tuesday, 05/Sept/2023 4:50pm - 5:10pm
ID: 207 / **S42: 3**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Machine Learning and Data Science
Keywords: dataset similarity, simulation studies

Comparison of methods for quantifying similarity of datasets

Marieke Stolte, Jörg Rahnenführer, Andrea Bommert
Department of Statistics, TU Dortmund University, Germany; stolte@statistik.tu-dortmund.de

Quantifying the similarity between two datasets has widespread applications in statistics and machine learning. Generalizability of a statistical model refers to the performance of the model on new or unseen datasets and depends on the similarity between the dataset used for fitting the model and the new datasets. In meta-learning and in transfer learning, a central component is to exploit or transfer insights between different datasets. In simulation studies, the similarity between distributions assumed in the simulation and distributions of datasets, for which the performance of methods is assessed, is crucial.

Extremely many approaches for quantifying dataset similarity have been proposed in the literature. Here, we present an extensive review and comparison of such methods. We examined more than 70 methods for quantifying the similarity of datasets and classified them into ten subclasses, including comparisons of cumulative distribution functions, comparisons of densities or characteristic functions, kernel- and graph-based discrepancy measures, methods based on inter-point distances, probability metrics, divergences, and comparisons based on binary classification. We compared all methods in terms of their applicability, interpretability, and theoretical properties, in order to provide recommendations for selecting an appropriate data similarity measure based on the goal of the dataset comparison and on the properties of the datasets at hand.

Based on insights from these comparisons, we aim to compare methods for simulating datasets (parametric, nonparametric, plasmode) to design more appropriate simulation studies. We will use the best-suited dataset similarity measures in a comparison of parametric and plasmode simulations for quantifying how similar simulated datasets are to data from the true data generating process. We present preliminary findings from a study for measuring the quality of regression models.

Thursday, 07/Sept/2023 11:20am - 11:40am
ID: 139 / **S66: 3**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Advanced survival analysis
Keywords: survival analysis, propensity score, bias

Consequences of omitted covariates on treatment estimates in propensity score matched studies

Alexandra Strobel[1], **Andreas Wienke**[1], **Oliver Kuß**[2]
[1]Institute of Medical Epidemiology, Biostatistics and Informatics, Interdisciplinary Center for Health Sciences, Medical Faculty of the Martin-Luther-University Halle-Wittenberg, Germany; [2]German Diabetes Center, Leibniz Center for Diabetes Research at Heinrich-Heine-University Düsseldorf, Institute for Biometrics and Epidemiology; Alexandra.strobel@uk-halle.de

Propensity score matching has become a popular method for estimating causal treatment effects in nonrandomized studies. However, for time-to-event outcomes, the estimation of hazard ratios based on propensity scores can be challenging if omitted or unobserved covariates are disregarded. Not accounting for such covariates could lead to heavily biased treatment estimates. Researchers often do not know whether (and, if so, which) covariates will induce this bias. To address this issue, we extended a previously described method, “Dynamic Landmarking”, which was originally developed for randomized trials. By simulation we show, that “Dynamic Landmarking” provides a good visual tool for detecting biased treatment estimates also in propensity score matched data. The underlying approach will be applied to a real world data set from cardiac surgery.

Monday, 04/Sept/2023 11:20am - 11:40am
ID: 116 / **S1: 2**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Neutral comparison studies in methodological research
Keywords: simulation studies, real data, meta learning

Against the "one method fits all data sets" philosophy for comparison studies in methodological research

Carolin Strobl[1], **Friedrich Leisch**[2]
[1]Universität Zürich, Schweiz; [2]Universität für Bodenkultur Wien, Österreich; carolin.strobl@uzh.ch

Many methodological comparison studies aim at identifying a single or a few "best performing" methods over a certain range of data sets. In this presentation we take a different viewpoint by asking whether the research question of identifying the best performing method is what we should be striving for in the first place. We will argue that this research question implies assumptions which we do not consider warranted in methodological research, that a different research question would be more informative, and how this research question can be fruitfully investigated.

Monday, 04/Sept/2023 11:20am - 11:40am
ID: 416 / S6: 2
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …)
Keywords: Variable selection; Model-based boosting; GAMLSS, Copula regression

Simplifying complex models: deselection for boosting distributional copula regression

Annika Strömer[1], Nadja Klein[2], Christian Staerk[1], Hannah Klinkhammer[1], Andreas Mayr[1]
[1]Department of Medical Biometrics, Informatics and Epidemiology, University Hospital Bonn, Germany; [2]Chair of Uncertainty Quantification and Statistical Learning, Research Center Trustworthy Data Science and Security (UA Ruhr) and Department of Statistics (Technische Universität Dortmund); stroemer@imbie.uni-bonn.de

Boosting distributional copula regression is a useful and flexible tool to jointly model multivariate outcomes, in which all parameters of the joint response distribution are related to covariates via additive predictors. Estimating and selecting the model through model-based boosting provides several useful features, such as the ability to model high-dimensional data situations. Additionally, boosting can incorporate data-driven variable selection simultaneously for all parameters of the marginal distributions as well as the association parameter of the copula. However, as known from univariate (distributional) regression models, the algorithm tends to select too many variables, particularly for low-dimensional settings ($p < n$). In these situations, the algorithm exhibits slow overfitting behaviour, resulting in the inclusion of many variables that have only minor importance and thus overall to a large model with difficult interpretation.

To counteract this behaviour, we adapt a recent deselection approach for statistical boosting to multivariate (copula) regression models to deselect base-learners with only a negligible impact on the overall performance of the model.

In a simulation study, we evaluate the performance of our deselection approach and additionally compare it to well-known methods to enhance variable selection such as stability selection and probing. All approaches effectively reduce the number of false positives. However, probing results in a lower predictive performance compared to the classical boosted model but with the smallest runtime. Stability selection and our deselection approach lead to a similar predictive performance as the classical approach whereas stability selection has the longest computational time. The latter renders stability selection infeasible for high-dimensional data.

Furthermore, we illustrate our deselection approach on high-dimensional genomic cohort data from the UK Biobank by modelling the joint genetic predisposition of two continuous phenotypes. Both outcomes are not only non-Gaussian distributed but also have an association that differs depending on the observed predictor variables, which justifies the need of a distributional copula regression model. Our results suggest that the approach is able to reduce the model complexity (improving therefore interpretability) and still leads to comparable results in terms of predictive performance.

Wednesday, 06/Sept/2023 11:20am - 11:40am
ID: 460 / **S51: 3**
Presentation Submissions - Featured Session
Featured Sessions: IBS-DR/ROeS Award session
Keywords: Random forest, Confounder adjustment, Genetic association studies, Model-based recursive partitioning algorithm, Maximally selected rank statistic

Confounder adjustment with random forests based on local residuals in genetic association studies

Annika Swenne
Leibniz Institute for Prevention Research and Epidemiology – BIPS, Germany; swenne@leibniz-bips.de

Random forest is a popular machine learning approach that is increasingly applied in genetic association studies. Genetic association studies aim at identifying genetic variants that are associated with a disease. This can be achieved, e.g. via the variable importance measures of the random forest. However, in the presence of confounders, disease-unrelated variants that are affected by a confounder could receive high importance scores, potentially masking associations with relevant variants. Zhao et al. [1] suggested to use the residuals from generalized linear models to adjust random forests for confounding. However, using the residuals from global linear models might not be sufficient if the influence of a confounder varies across unknown subgroups.

In this work, I use the model-based recursive partitioning algorithm [2] in the construction of the random forest trees to automatically identify these subgroups. Since the model-based recursive partitioning algorithm uses computationally expensive generalized M-fluctuation tests to determine the splits of the random forest trees, it might not be feasible for large genetic association studies. Therefore, I develop and compare two modifications of the algorithm that use the residuals from local linear models to determine the splits: the residual variance splitting and the maximally selected residual rank statistic splitting rule. The first modification is similar to the original CART algorithm [3], whereas the latter is based on the maximally selected rank statistic [4].

The results of my simulation studies show that only the maximally selected residual rank statistic splitting rule is able to adjust for confounding. Further, this method performs slightly better than the global adjustment if the confounder model only holds for subgroups of the population. On real data from a European cohort study on child health, the proposed method leads to a sparser solution, i.e., it selects fewer variants, but still identifies those that are known to be associated with the outcome.

References

[1] Zhao, Y. et al. (2012). ”Correction for Population Stratification in Random Forest Analysis”. In: International Journal of Epidemiology 41.6, pp. 1798–1806. doi: 10.1093/ije/dys183.

[2] Zeileis, A., T. Hothorn, and K. Hornik (2008). ”Model-Based Recursive Partitioning”. In: Journal of Computational and Graphical Statistics 17.2, pp. 492–514. doi: 10.1198/106186008X319331.

[3] Breiman, L. et al. (1984). Classification and Regression Trees. New York: Chapman & Hall/CRC. doi: 10.1201/9781315139470.

[4] Lausen, B. and M. Schumacher (1992). ”Maximally Selected Rank Statistics”. In: Biometrics 48.1, pp. 73–85. doi: 10.2307/2532740.

ID: 463 / **Plenary 2: 1**
Presentation Submissions - Featured Session
Featured Sessions: Keynote
Keywords: Microbiome data analysis, Negative binomial hurdle model, Alpha diversity, Beta diversity, Bacterial community composition

Statistical methods to analyse the structure of the microbiome based on cereal leaf beetle (Oulema melanopus) data

Alicja Szabelska-Beręsewicz[1], Beata Wielkopolan[2], Krzysztof Krawczyk[2], Aleksandra Obrępalska-Stęplowska[2]
[1]Poznań University of Life Sciences, Poland; [2]Institute of Plant Protection – National Research Institute, Poland;
aszab@up.poznan.pl

Insects are an integral part of the biodiversity in virtually all terrestrial ecosystems, making them very important for environmental impact assessment. The cereal leaf beetle (CLB, *Oulema melanopus*) is a serious agricultural pest that causes significant damage to agricultural production. Recent advances in next-generation sequencing (NGS) have made it possible to study microbial communities with unprecedented resolution. However, these advances in data generation have created new challenges for researchers attempting to analyse and visualise these data.

The aim of this talk is to present statistical methods to characterise the bacterial communities associated with CLB larvae and imagoes collected from different cereal host species and locations. Identification of CLB-associated bacteria was performed by 16S rRNA gene sequencing at the V3-V4 hypervariable region.

To test which factors influence the microbial composition at each taxonomic level, the abundance of the bacterial community can be modelled with a negative binomial hurdle model (hNB GLMM). Methods for checking the appropriateness of the choice of hurdle model, evaluating parameter estimation and assessing fit of the model will be presented. In addition, permutational procedures for assessing the importance of variables will be discussed. Next, the term top biome will be introduced to provide information on which taxa have an abundance that is significantly different from the overall mean abundance of all taxa. Methods for determining of bacterial biodiversity will be presented. Different identifiers for alpha and beta biodiversity will be discussed.

The methods presented can contribute to a better understanding of the microbiome composition, diversity, and the main factors influencing the content of insect-associated bacteria, and can be successfully applied in other ecological studies.

Monday, 04/Sept/2023 2:20pm - 2:40pm
ID: 288 / **S13: 2**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), Software Engineering, Time-to-Event Analysis
Keywords: additive models, mixed-effects models, R package, regression, transformation models

Mixed-effects Additive Transformation Models with the R Package tramME

Balint Tamasi
University of Zurich, EBPI, Switzerland; balint.tamasi@uzh.ch

Regression models that accommodate correlated observations and potential nonlinearities in the relationship between predictors and outcome are important tools in the analysis of experimental and observational data. Traditional parametric approaches assume that the conditional response distribution can be fully captured with a few parameters of a predefined distribution. In practice, finding the correct distribution type can be problematic, and misspecifications may lead to inefficient or incorrect inference. Transformation models approximate the conditional outcome distribution in a data-driven way, using flexible parameterizations, which makes the approach universally applicable to any, at least ordered, outcome types. This talk presents an extension of the transformation model framework with general random effect structures and penalized smooth terms, to adapt it to practical settings with complex correlated data and nonlinear predictor-outcome relationships. The R package tramME provides an implementation of the methodology by connecting functionalities from popular and well-tested R packages for transformation modeling (mlt), mixed-effects (lme4) and additive models (mgcv). With the Template Model Builder (TMB) framework in its computational core, the package provides fast and efficient likelihood-based estimation and inference in the general class of mixed-effects additive transformation models. The usage of tramME is demonstrated through an analysis of an ecological experiment with interval-censored and grouped time-to-event responses and an individual participant data meta-analysis of a collection of observational studies about burn patient recovery with bounded quality-of-life outcomes.

Thursday, 07/Sept/2023 9:10am - 9:30am
ID: 242 / **S61: 2**
Presentation Submissions - Invited Session
Invited Sessions: Statistical strategies in toxicology
Keywords: DILI, Toxicity, Prediction, Liver Injury, Modeling

An integrated data-driven approach for drug safety prediction

Fetene Tekle[1], **Vahid Nassiri**[2], **Kanaka Tatikola**[1], **Helena Geys**[1]
[1]Janssen R&D, Belgium; [2]Open Analytics, Belgium; ftekle@its.jnj.com

Predicting drug-induced organ injury is a multi-dimensional problem that requires consideration of multifaceted assays and additional compound-related information. Chemical properties of compounds, together with multiple in-vitro assays endpoints and exposure parameters such as dose, are promising markers to detect drug-induced organ injury in early drug development. Tools that can help to integrate data from different sources can make the decision-making process faster, data driven, and more efficient. In this talk, I will describe one of such tools for predicting drug-induced liver injury (DILI) and discuss how such approaches can also help predict other organ injuries. The proposed methodology will be validated by implementing a statistical model on three publicly available datasets. Model based probabilities will be used to classify the predicted DILI risk class corresponding to the true DILI severity. Classification methods such as Winner's rule (maximum probability), LDA (linear discriminant analysis) and QDA (quadratic discriminant analysis) will be discussed and compared.

Thursday, 07/Sept/2023 11:00am - 11:20am
ID: 287 / **S69: 2**
Presentation Submissions - Regular Session (Default)
Topic of Submission: High dimensional data, genetic and x-omics data
Keywords: variable selection, high dimensional data, cross leverage scores, interaction effects

Detecting interactions in High Dimensional Data using Cross Leverage Scores

Sven Teschke[1,2], Katja Ickstadt[1], Alexander Munteanu[1], Tamara Schikowski[2]
[1]TU Dortmund, Germany; [2]IUF Düsseldorf, Germany; sven.teschke@tu-dortmund.de

We are developing a variable selection method for regression models in Big Data in the context of Genetics. In particular, we want to detect important interactions between variables. The method is intended for investigating the influence of SNPs and their interactions on health outcomes, which is a $p \gg n$ problem.

Motivated by Parry et al. (2021), we use the so called cross leverage scores to detect interactions of variables while maintaining interpretability. The big advantage is that this method does not require considering each possible interaction between variables individually, which would be very time consuming even for a moderately large amount of data. In a simulation study we show that these cross leverage scores are directly linked to the importance of a variable in the sense of an interaction effect.

Furthermore, we are developing methods for the detection of interactions using cross leverage scores to very large datasets, as they are common in the context of genetics. The key idea is to divide the data set into subsets of variables (batches). Successively, for each batch we store the (predefined) q most important variables, compare them to those selected from the previous batch, store the combined q most important variables and reject the rest. We receive the q most important variables of the whole data set after analyzing all batches. Thus, we avoid complex and time-consuming computations of high-dimensional matrices by performing the computations only for small batches of the partitioned data set, which is much less costly. We then also compare these methods to existing approximation methods for calculating cross leverage scores (Drineas et al. (2012)). We evaluate these methods with simulation studies and with a real data set, the SALIA study (Study on the Influence of Air Pollution on Lung, Inflammation and Aging) (Schikowski et al. (2005)). This study investigates the influence of air pollution on lung function, inflammatory responses and aging processes in elderly women from the Ruhr area. Since we are particularly interested in genetic data, we consider $n=517$ women from this study. In addition to data on influences of various environmental factors, data on over 7 million SNPs is also available for these women. We are exploring the influence of both SNP interactions and SNP environment interactions on various health outcomes.

References:

1. Drineas P., Magdon-Ismail M., Mahoney M.W., Woodruff D.P. (2012). Fast approximation of matrix coherence and statistical leverage. J Machine Learning Research 13, 3475-3506, doi: 10.5555/2503308.2503352.
2. Parry, K., Geppert, L., Munteanu, A., Ickstadt, K. (2021). Cross-Leverage Scores for Selecting Subsets of Explanatory Variables. arXiv e-prints, abs/2109.08399, https://arxiv.org/abs/2109.08399.
3. Schikowski, T., Sugiri, D., Ranft, U. *et al.* (2005). Long-term air pollution exposure and living close to busy roads are associated with COPD in women. *Respir Res* **6**, 152. https://doi.org/10.1186/1465-9921-6-152

Tuesday, 05/Sept/2023 4:10pm - 4:30pm
ID: 359 / **S36: 1**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Clinical trial designs (adaptive and platform designs, external controls, …), Statistical hypothesis testing (covariate adjustment, nonparametric methods, multiple comparisons, …)

Practical learnings from futility analyses in Phase 3 trials

Gian-Andrea Thanei, Gaëlle Klingelschmitt, Claude Berge, Hans-Ulrich Burger
Roche, Switzerland; gian-andrea.thanei@roche.com, gaelle.klingelschmitt@roche.com, claude.berge@roche.com, hans_ulrich.burger@roche.com

Futility analysis is a critical component of clinical trials and should always be part of the discussion when setting up a clinical trial. It requires clear communication from statisticians to all stakeholders to make informed decisions. In this presentation, we share our recent experience with setting up a futility analysis and the practical learnings we gained from it. Specifically, we discuss the key risks involved in futility analysis and how we effectively communicated these risks to decision makers. Additionally, we propose a set of metrics and visualizations that can help facilitate discussions with non-statistical stakeholders.

Tuesday, 05/Sept/2023 12:20pm - 12:40pm
ID: 254 / **S22: 4**
Presentation Submissions - Invited Session
Invited Sessions: Net benefit, win odds, and win ratio: Methods, analysis, and interpretation
Keywords: GPC, Probabilistic Index Models, Semiparametric Inference

Semiparametric and Nonparametric Methods for Covariate Adjustment for GPC Effect Sizes for Multiple Outcomes

Olivier Thas
Data Science Institute, I-BioStat, Hasselt University, Belgium; olivier.thas@UHasselt.be

Generalised pairwise comparisons (GPC) is gaining more and more attention as an effect size in clinical trials. It can take several forms (e.g net benefit, win ratio, win odds, probabilistic index) and can be defined for a single outcome as well as for multiple outcomes. The estimation of this effect size, and its properties, is still an ongoing research area, and correcting the GPC effect size for covariates has not yet attracted much attention.

We have developed flexible semiparametric methods for analysing the net benefit with adjustment for baseline covariates. These methods are based on Probabilistic Index Models (PIM) and influence functions under nonparametric models, and they are easy to implement. In this talk, we will outline the construction of the methods and demonstrate them in a case study.

Monday, 04/Sept/2023 11:40am - 12:00pm
ID: 233 / **S7: 3**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical hypothesis testing (covariate adjustment, nonparametric methods, multiple comparisons, …)
Keywords: ANCOVA, nonparametrics

Robust ANCOVA for Small Sample Studies

Konstantin Emil Thiel[1,2], **Georg Zimmermann**[1], **Arne C. Bathke**[2]
[1]Paracelsus Medical University Salzburg, Austria; [2]University of Salzburg, Austria; konstantin.thiel@pmu.ac.at

Evaluating the effect of a group variable on a response variable, while controlling for nuisance variables (covariates), is referred to as analysis of covariance (ANCOVA). The idea is to reduce response variance by accounting some of the total variance to the covariates, and thus, increase power of statistical tests for group effects. Besides theory for parametric ANCOVA, which is restricted to normally distributed data, there exist also some semi- and nonparametric approaches. However, none of the existing approaches can be reliably applied to small sample studies, since various simulations have indicated poor type-I error control in these situations. We close this gap by enhancing existing methods with newly developed small sample tests. In other words, we build a robust ANCOVA that is applicable when classical parametric assumptions are violated or when available data is limited. The performance of our tests is evaluated in extensive simulations that demonstrate an improved type-I error control, while maintaining competitive power.

Monday, 04/Sept/2023 2:40pm - 3:00pm
ID: 174 / **S14: 3**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Clinical trial designs (adaptive and platform designs, external controls, …), Statistical hypothesis testing (covariate adjustment, nonparametric methods, multiple comparisons, …)
Keywords: co-primary endpoints, dual primary endpoints, at-least-one concept, trial design, confirmatory clinical trials

Dual Primary Endpoints – innovative idea or avoidable risk?

Nele Henrike Thomas, Armin Koch, Anika Großhennig
Medical School Hannover, Germany; thomas.nele@mh-hannover.de
Note: A companion presentation to this contribution will be given at the "68. GMDS-JAHRESTAGUNG 2023".

Introduction: Formal proof of efficacy of a new drug requires that in a prospective clinical trial, superiority towards a placebo or an established standard is demonstrated. Traditionally one primary endpoint is specified, but several diseases (e.g., Alzheimer's disease, cancer) exist where treatment can be based on the assessment of multiple primary endpoints. With two (or more) co-primary endpoints, significant superiority must be shown for both to claim study success. No adjustment of the study-wise type-1-error is needed in this scenario, but the sample size usually needs to be increased to maintain the pre-defined power. Recently, also trials that use the so-called dual primary endpoint concept have been proposed. Here, a Bonferroni split is applied to guarantee that the study-wise type-1-error is controlled at the pre-specified level. Study success is claimed as soon as superiority for at least one of the primary endpoints is demonstrated. This approach is sometimes also called the at-least-one concept. In contrast to co-primary endpoints, it is not covered in the European guideline on multiplicity because study success can be declared as soon as superiority for one primary endpoint is demonstrated, even if the other indicates a deterioration. This is obviously illogical and not in line with practical decision-making and thus may lead to post-hoc discussions on overall study success if results for one of the endpoints are ambiguous.

Methods: We investigate the situation with two primary endpoints where it may be sufficient to demonstrate superiority in at least one of them. In line with Röhmel et al., we argue that interpretational issues should be discussed upfront and examine several additional constraints to the dual primary endpoint concept to assure a statistically and clinically consistent strategy for decision-making. Our principal idea is first to require a certain "minimal requirement" for all primary endpoints before addressing the superiority hypothesis. Specifically, either the treatment effect estimate must be on the right side of zero, a positive trend is required for the primary endpoints, or non-inferiority to a pre-defined margin has to be demonstrated. Additionally, superiority to the control has to be shown for at least one of the primary endpoints. We performed a simulation study to examine our approach. The main intention was to compare our decision strategy to the dual primary endpoint concept regarding increased costs (in terms of power (and sample size)). Simulation scenarios included various treatment effects and two correlations for three different sample sizes.

Results/Conclusion: Our simulations illustrate that if the treatment effects are as planned, additional constraints to the dual primary endpoint concept lead to statistically and clinically consistent decisions on study success and improve interpretation with limited costs in terms of sample size. Moreover, our approach allows flexible modeling of the minimum requirements for all endpoints and leads back to the co-primary endpoint concept reflected in the European multiplicity guideline. In summary, our work emphasizes that the more aspects are discussed and pre-defined at the planning stage of a clinical trial, the better the certainty and interpretability of its results.

Tuesday, 05/Sept/2023 5:10pm - 5:30pm
ID: 193 / S42: 4
Presentation Submissions - Regular Session (Default)
Topic of Submission: Time-to-Event Analysis

How to Simulate Realistic Survival Data? A Simulation Study to Compare Realistic Simulation Models

Maria Thurow[1], Ina Dormuth[1], Christina Nießl[2], Anne-Laure Boulesteix[2], Marc Ditzhaus[3], Markus Pauly[1,4]
[1]TU Dortmund University, Germany; [2]Ludwig-Maximilians-University of Munich, Germany; [3]Otto-von-Guericke-University Magdeburg, Germany; [4]UA Ruhr, Research Center Trustworthy Data Science and Security, Dortmund, Germany;
maria.thurow@tu-dortmund.de

In statistics, it is important to have realistic data sets available for a particular context to allow an appropriate and objective method comparison. For many use cases, benchmark data sets for method comparison are already available online. However, in most medical applications and especially for clinical trials in oncology, there is a lack of adequate benchmark data sets, as patient data can be sensitive and therefore cannot be published. Another possible challenge is the need of a larger number of data sets or observations. Furthermore, if methods need to be compared for a specific setting, this may not be covered by the available data sets. A potential solution for this are simulation studies.

However, it is sometimes not clear, which simulation models are suitable for generating realistic data, e.g., for time-to-event analyses. A challenge here is that potentially unrealistic assumptions have to be made about the distributions. Instead, benchmark data sets can be used as a basis for the simulations, which has the following advantages: the actual properties are known and more realistic data can be simulated. There are several possibilities to simulate realistic data from benchmark data sets. In order to make recommendations on which models are best suited for a specific survival setting, we conducted a simulation study comparing simulation models based upon kernel density estimation, fitted distributions, and two different bootstrap approaches. We used the runtime and different accuracy measures (e.g., the p-values of the log-rank test wrt the benchmark data sets) as criteria for comparison.

Using the example of comparing two-sample procedures for lung cancer studies, we propose a way to simulate realistic survival data in two steps: In a first step, we provide reconstructed benchmark data sets from recent studies on lung cancer patients. To do so, we first searched for adequate data sets. Therefore, we considered phase III clinical studies from oncology in which a log-rank test (our benchmark method) was applied and the Kaplan-Meier estimator was reported. This resulted in 290 potential studies. Restricting our analysis to lung cancer as a gender-independent cancer type and to studies with the required necessities for reconstruction, finally resulted in seven studies. We then reconstructed the data sets using a state-of-the-art reconstruction algorithm. In a second step, we build upon the reconstructed benchmark data sets to propose different realistic simulation models for model comparison.

Besides slight differences in runtime, our results show that, in our setting, simulations based on kernel density estimation and case resampling lead to data sets representing the original data well while simulating data from a fitted distribution does not succeed to do so. This demonstrates that it is possible to simulate realistic survival data when benchmark data sets (or at least the required information to reconstruct them) from real-world studies are available. In subsequent future applications, these results can be used for method comparison or further analyses, e.g., for sample size planning (for follow-up studies).

Tuesday, 05/Sept/2023 11:00am - 11:20am
ID: 421 / S24: 1
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), Machine Learning and Data Science
Keywords: Bayesian Inference, Synthetic Data Generation, Uncertainty Quantification, Robustness, Deep Learning

Bayesian Uncertainty Quantification in Deep Generative Models for Synthesis of Tabular Medical Data

Patric Tippmann, Kiana Farhadyar, Daniela Zöller
Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center – University of Freiburg, Germany;
patric.tippmann@uniklinik-freiburg.de

Medical research and patient care rely on the collection, analysis, and reuse of medical data. However, data access is often limited (e.g., due to data protection constraints). Synthetic data generation is a promising solution to enable researchers to generate synthetic medical data that preserves the statistical properties of the original data while ensuring privacy. However, interpreting the results using synthetic data needs to account for uncertainty in the generation process, especially when using deep generative models. We aim to address this problem by employing Bayesian inference techniques for uncertainty quantification.

Specifically, we focus on Variational Autoencoders (VAEs) to generate tabular medical data since they provide a probabilistic framework by explicitly modeling the probability distribution of the data while simultaneously providing a latent low-dimensional data space for additional investigations. However, the overconfidence of deep neural networks (DNNs) for anomalous or Out-of-Distribution (OOD) data is an unsolved problem, which can lead to unreliable model predictions or inflated probability estimates in the downstream analysis of the synthetic data. Reliable methods for quantifying uncertainty are necessary to address this issue, yet previous work has mostly addressed it in the context of supervised discriminative models. Very recent literature also explores Bayesian methods for uncertainty estimation in VAEs, albeit in the context of image data. Tabular medical data impose greater challenges as they often contain missing or anomalous values, which can introduce bias or lead to inaccurate models if not handled properly. Moreover, such datasets have heterogeneity in the data distributions (e.g., complex and multi-modal) and data types (e.g., discrete and continuous variables) that require specialized treatment.

In our approach, we address these challenges with a suitably tailored VAE framework and compare two Bayesian inference methods to quantify epistemic (knowledge-related) model uncertainty based on model averaging and Markov-Chain Monte Carlo techniques. We review and apply robust metrics to evaluate the quality of the generated data, showing that our approach preserves important properties of the original data. Based on both real and simulation data, we demonstrate how the proposed approach improves the faithfulness of downstream task performance, such as classification and regression, by providing more accurate and reliable synthetic data.

In summary, our work covers the use of Bayesian inference for uncertainty quantification in a new area, namely synthesizing tabular medical data using VAEs. We highlight the challenges and considerations that arise with tabular synthetic data generation. Using real and simulated medical datasets, we show how our approach and framework can increase model usefulness. Our work contributes towards developing more reliable deep generative models for medical applications.

Thursday, 07/Sept/2023 9:50am - 10:10am
ID: 327 / **S61: 4**
Presentation Submissions - Invited Session
Invited Sessions: Statistical strategies in toxicology
Keywords: in vivo alkaline comet assay, hierarchical mixed effect models, handling of zeros, interaction of negative and positive controls, influence of the slide summary

THE COMET ASSAY IN VIVO – A REVIEW OF KNOWN PROPERTIES AND NEW FINDINGS

Timur Tug[1], Julia Duda[1], Bernd-Wolfgang Igl[2], Katja Ickstadt[1]
[1]Department of Statistics, TU Dortmund University, Dortmund, Germany; [2]2Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach an der Riss, Germany; timur.tug@tu-dortmund.de

The *in vivo* alkaline comet assay is a widely used and regulatory-relevant test in genotoxicity. It is a sensitive and fast method to detect both, DNA strand break induction, and DNA repair on a single cell level.

After a brief description of the biological background and the technical performance of the *in vivo* test, several fundamental statistical aspects of Comet data will be examined. This includes the description of the data distribution, the handling of zeros, the interaction of negative and positive controls, the inclusion of historical negative control data and the influence of the chosen slide summary on the final test outcome.

Based on a large data set, we were able to validate the OECD 489 suggested treatment of zero values and to endorse the median as the proposed summary measure. We also offer advice on how to compare negative and positive controls for assessing the validity of a study. Moreover, a variance decomposition analysis offers insightful information on the origin of noise on the cell, slide, and animal level, which may influence the experimental design.

For this purpose, simple hierarchical mixed effect models were set up and in addition, more complex models were used to improve the understanding of the structure of the comet assay. Nevertheless, open points will be sketched to be considered in the future.

Thursday, 07/Sept/2023 8:30am - 8:50am
ID: 352 / **S60: 1**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …)
Keywords: variable selection, regression, prediction

Do we need different variable selection procedures depending on the goal of the statistical model?

Theresa Ullmann, Georg Heinze, Daniela Dunkler
Medical University of Vienna, Center for Medical Data Science, Section for Clinical Biometrics, Vienna, Austria;
theresa.ullmann@meduniwien.ac.at

Data-driven variable selection is frequently performed in statistical modeling, i.e., when modeling the associations between an outcome and multiple independent variables. Variable selection may improve the interpretability, parsimony and/or predictive accuracy of a model (Heinze et al. 2018, doi:10.1002/bimj.201700067). Many different methods for variable selection have been proposed, such as selection based on significance criteria (e.g., backward elimination), or methods based on penalized likelihoods (e.g., the LASSO).

Less attention has been given to the fact that the specific purpose of variable selection depends on the *goal* of modeling. Shmueli (2010, doi:10.1214/10-STS330) distinguished between three main types of statistical models: descriptive, explanatory, and predictive. In *descriptive* modeling, researchers aim to describe the relationship between the outcome and the independent variables in a parsimonious manner. Here, variable selection may help to generate simple and interpretable models. In *explanatory* modeling, researchers are interested in estimating the causal effect of a specific explanatory variable, often an intervention, on the outcome adjusted for confounders. The confounders are typically chosen a priori based on domain expertise (e.g., with the help of directed acyclic graphs). Still, researchers might expect data-driven variable selection to increase the precision of the effect estimate by eliminating confounders with negligible association with the outcome. Finally, in *predictive* modeling, the main goal is to predict the outcome as accurately as possible. Here, variable selection may help to remove noise and thus reduce the prediction error. Sometimes, modeling has multiple goals, e.g., to find a descriptive model that is also suitable for prediction. In such situations, variable selection must serve multiple purposes.

In this talk, we will first discuss variable selection in the context of different modeling goals. Then we will present the results of a simulation study where we evaluated different variable selection methods, including backward elimination, the LASSO, and others. Multivariable data is simulated based on real-world data from the National Health and Nutrition Examination Survey (NHANES). Different sample sizes and R^2 are considered. We evaluate the results according to various performance criteria (e.g., the effect of the selection on the bias and variance of the coefficient estimation, the effect on the prediction error, as well as the selection rate of the 'true' model). In the interpretation of the results, we put a particular focus on which estimands and performance criteria are most relevant for which modeling goals. For example, in explanatory modeling, the effect estimator is the main estimand, whereas in descriptive modeling, a particular focus is on selecting the 'true' model, or at least not missing the most relevant variables. For any method, there is a strong association of any type of performance with sample size and the underlying R^2. Consequently, the choice of a variable selection method should take into account knowledge and assumptions about these main drivers of performance, but also the modeling goal. Our talk encourages data analysts to think carefully about their modeling goals before planning their modeling analysis.

This work was supported through the Austrian Science Fund FWF [project I-4739-B].

Wednesday, 06/Sept/2023 11:20am - 11:40am
ID: 249 / S55: 3
Presentation Submissions - Regular Session (Default)
Topic of Submission: Preclinical drug development, safety and toxicology
Keywords: Historical Control Data, Virtual Control Groups, Systemic Toxicity Study, Replacement, Clinical Chemistry

Hurdles and Signposts on the Road to Virtual Control Groups in Toxicity Studies

Lea A.I. Vaas[1], Alexander Gurjanov[2], Guillemette Duchateau-Nguyen[3], Annika Kreuchwig[2], Hannes-Friedrich Ulbrich[1], Frank Bringezu[4], Matteo Piraino[5], Thomas Steger-Hartmann[2], eTRANSAFE Consortium[6]

[1]Bayer AG, Research & Development, Pharmaceuticals, Research & Pre-Clinical Statistics; [2]Bayer AG, Research & Development, Pharmaceuticals, Investigational Toxicology; [3]Roche Pharmaceutical Research and Early Development, Pharmaceutical Sciences Roche Innovation Center Basel, Switzerland; [4]Merck Healthcare KGaA, Biopharma, Chemical & Preclinical Safety; [5]Data Lab for Research and Innovation, Organon SRL, Bucharest, Romania; [6]www.etransafe.eu/partners-etransafe/; lea.vaas@bayer.com

In systemic toxicity studies replacement of concurrent control animals by so-called Virtual Control Groups (VCGs) may reduce the use of animals thus contributing to the 3R's principle of animal experimentation: Replacement, Reduction, and Refinement.

Within the Innovative Medicine Initiatives project eTRANSAFE (**e**nhancing **TRANS**lational **SAFE**ty Assessment through Integrative Knowledge Management)[1] the VCG subgroup was formed tackling the current major obstacles: (i) collection, curation and sharing of suitable sets of historical control data from preclinical toxicity studies, (ii) investigation of methodologies for derivation of ViCoGs from historical control data and performance testing in statistical analysis and (iii) to reach out for regulatory advice to gain acceptance of this concept as early as possible[2].

This talk introduces the ViCoG database with currently (FEB 2023) more than 60 000 animals' data donated from Bayer AG, Merck KGaA, Novartis Pharma AG, F. Hoffmann - La Roche AG, Sanofi Aventis GmbH and Fraunhofer Gesellschaft ITEM. The SEND (Standard for Exchange of Non-clinical Data)[3] domains Demographics, Organ Measurements, Clinical Observations, Macroscopic Findings, Body Weights, Laboratory Measurements and Microscopic Findings are covered and are populated with between >65.000 and up to >257.000 records per domain. All partners contribute with their expertise to this unique cross-industry resource for advancing both VCG-method development, data reuse and rethinking statistical analysis frameworks in toxicity studies in general.

A proof of concept based on 67 toxicology four-week studies revealed overall good agreement between original test results and reanalysis using VCGs for 19 clinical chemistry parameters.

Emphasizing SEND regarding data-formatting and provision of required experimental information, a case study on the effect of a hidden confounder illustrates the special role of understanding sources of variability within the data[4].

Strategies to address the impact of the hidden confounder are presented leading ultimately to a proposal of dedicated control-chart approaches fostering provision of required experimental information for fine-tuned data-selection in a VCG-framework.

In an outlook the impact of FDA's recent changes allowing promotion of drug-candidates to human trials after either animal or nonanimal tests, on the potential role of VCGs in nonanimal testing situations will be considered.

References:

[1] www.etransafe.eu/the-project/about/

[2] Steger-Hartmann, T., Kreuchwig, A., Vaas, L., Wichard, J., Bringezu, F., Amberg, A., Muster, W., Pognan, F. and Barber, C. (2020) Introducing the concept of virtual control groups into preclinical toxicology testing, ALTEX - Alternatives to animal experimentation, 37(3), pp. 343–349. doi: 10.14573/altex.2001311.

[3] www.cdisc.org/standards/foundational/send

[4] Gurjanov, A., Steger-Hartmann, T., Kreuchwig, A., and Vaas, L.A.I.(2023) Hurdles and Signposts on the Road to Virtual Control Groups -A Case study illustrating the Influence of Anesthesia Protocols on Electrolyte Levels in Rats, Front. Pharmacol. (under review)

Monday, 04/Sept/2023 5:30pm - 5:50pm
ID: 328 / **S17: 5**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Prognostic and predictive biomarkers in personalized medicine
Keywords: treatment effect modification, interaction, study design, subgroup analyses

Searching for treatment effect modifiers in manual therapy: Three case studies

Werner Vach
Basel Academy for Quality and Research in Medicine, Switzerland; werner.vach@basel-academy.ch

For many disorders a musculoskeletal problem is a potential explanation and manual therapy may be a promising treatment option. However, it is often not possible to identify the musculoskeletal problem and for many patients it remains only one of several potential explanations. It has to be expected that manual therapy is only beneficial for those patients with a musculoskeletal problem, and hence it is of interest to identify patient characteristics which may be associated with the presence of a musculoskeletal problem and hence may be predictive for the treatment benefit from manual therapy. This allows to use domain knowledge and domain expertise to identify promising treatment effect modifiers even prior to the first RCT on manual therapy for a specific disorder. As domain knowledge and domain expertise is used, it is also possible to use hypotheses about the direction of the modification to specify a priori a summary index to predict the treatment effect. This process of identifying promising effect modifiers is exemplified using two RCTs on the effect of manual therapy on disorders potentially caused by musculoskeletal problems: colic in infants (Holm et al 2021) and recurrent headache in children (Lynge et al 2021). Both RCTs are analysed using the same strategy, combing a confirmatory part with an exploratory part. The strategy of identifying the effect modifiers is related to the success of the analysis to find indications for effect modification. A third example focus on the identification of interactions between the treatment setting (chiropractice vs general practice) and the clinical outcome based on observational data (Hartvigsen et al 2020). In this example, interactions are not interpreted as treatment effect modifications, but as differences in the prognostic value of patient characteristics between different settings. Challenges in an interpretation of this type are discussed.

References:

1. Hartvigsen L, Kongsted A, Vach W, Salmi LR, Hestbaek L (2020). Baseline Characteristics May Help Indicate the Best Choice of Health Care Provider for Back Pain Patients in Primary Care: Results From a Prospective Cohort Study. J Manipulative Physiol Ther. 43(1):13-23. doi: 10.1016/j.jmpt.2019.11.001
2. Holm LV, Vach W, Jarbøl DE, Christensen HW, Søndergaard J, Hestbæk L (2021). Identifying potential treatment effect modifiers of the effectiveness of chiropractic care to infants with colic through prespecified secondary analyses of a randomised controlled trial. Chiropr Man Therap. 29(1):16. doi: 10.1186/s12998-021-00373-6.
3. Lynge S, Dissing KB, Vach W, Christensen HW, Hestbaek L (2021). Effectiveness of chiropractic manipulation versus sham manipulation for recurrent headaches in children aged 7-14 years - a randomised clinical trial. Chiropr Man Therap 29(1):1. doi: 10.1186/s12998-020-00360-3.

Monday, 04/Sept/2023 2:40pm - 3:00pm
ID: 316 / **S11: 3**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Safety and benefit/risk assessment in master protocols
Keywords: benefit-risk assessment; master protocols; safety.

Safety and Benefit-Risk Evaluation: Master Protocols for Efficient Evidence Generation

Alessandro Vagheggini
Clinical Safety Statistics - Biostatistics and Research Decision Sciences MSD Innovation and Development GmbH, Switzerland; alessandro.vagheggini@msd.com

Master protocols are becoming increasingly utilized as the clinical trial designs of choice for the new precision medicine paradigm. This can come in the form of master protocols with the same experimental therapies in cohorts of different disease populations (i.e., basket trials) or multiple experimental arms with a shared control arm (i.e., umbrella trials). Many adaptive features are available to tailor the design for efficient evidence generation. As customary, the bulk of these methods focus solely on efficacy providing methodological tools able to identify the most promising treatments and/or populations. At the same time, safety is usually investigated by means of descriptive incidences rather than using more advanced methods.

An overarching goal of any drug development program is to evaluate benefit-risk profile of pharmaceutical products. Our research will focus on (1) how to leverage information across sub-components – e.g., multi-cohorts or multi-arms – of the master protocol for efficient aggregate safety evaluation and (2) how to systematically combine safety and efficacy evaluations in master protocols by means of structured benefit-risk assessment (BRA) tools. Key concepts will be introduced to highlight different considerations to be made through the various phases of any drug development lifecycle. Available qualitative and quantitative approaches to BRA will also be outlined.

Thursday, 07/Sept/2023 11:20am - 11:40am
ID: 114 / **S65: 2**
Presentation Submissions - Invited Session
Invited Sessions: Advancing clinical trial design in rare diseases
Keywords: hybrid design, clinical trials, amyotrophic lateral sclerosis

Hybrid controlled clinical trials using concurrent registries in Amyotrophic Lateral Sclerosis: A feasibility study

Ruben van Eijk
UMC Utrecht, Netherlands, The; r.p.a.vaneijk-2@umcutrecht.nl

Background: Hybrid designs with a randomised and external control arm preserve key features of randomization and utilize patient-level information from registries and electronic health records to augment clinical trials. Integration of external control data, however, can distort study results if the external cohort is not interchangeable with the randomised controls. In this study, we propose to leverage high-quality, patient-level concurrent registries for augmenting clinical trials and illustrate its impact on trial design for Amyotrophic Lateral Sclerosis (ALS).

Methods: The proposed methodology was evaluated in a completed, randomised, placebo-controlled clinical trial. We used patient-level information from a well-defined, population-based registry that ran parallel to the clinical trial to reconstruct the eligible trial population, identify concurrently non-participating eligible patients who matched with trial participants, and integrate them into the statistical analysis. We assessed the impact of the matched external data on the treatment effect estimate, precision, and time to reach a conclusion.

Findings: During the runtime of the trial, a total of 1,141 patients were alive in the registry, of whom 473 (41·5%) fulfilled the eligibility criteria and 133 (11·7%) participated. Trial participants were younger and in a better health compared to non-participating eligible patients. A matched control population could be identified among the non-participating patients. Integration of external controls into a hybrid design improved precision and increased statistical power from 58·1% to 77·3%. Augmenting the randomised controls with matched external controls could prevent randomization of an additional 17 patients (12·8%) and reduce the trial duration by 25·0% to terminate a futile trial. Matching eligible external controls from a different calendar period did not lead to improvement.

Interpretation: Hybrid trial designs utilizing a concurrent patient-level registry with rigorous matching may minimise bias due to a mismatch in calendar time and differences in care pathways, and can accelerate the development of new treatments.

Wednesday, 06/Sept/2023 10:40am - 11:20am
ID: 173 / S50: 1
Presentation Submissions - Invited Session
Invited Sessions: Covariate Adjustment in RCTs: Translating theory into practice within a pharmaceutical company via a data challenge
Keywords: covariate adjustment, causal inference, standardization, treatment policy, robustness

Improving Power in Randomized Trials by Leveraging Baseline Variables

Kelly Van Lancker
Ghent University, Belgium; kelly.vanlancker@ugent.be

In many clinical trials, data is collected on different patient characteristics at the time of entry (e.g., age, baseline severity and comorbidities). Covariate adjusted estimation methods that can both be more efficient than unadjusted estimators whilst also remaining robust to model misspecification (i.e., we require consistent estimators user arbitrary model misspecification) are available. The resulting sample size reductions can lead to substantial cost savings, and also can lead to more ethical trials since they avoid exposing more participants than necessary to experimental treatments. This was also emphasized in the recent guidance released by the U.S. Food and Drug Administration (FDA) for industry on “Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products”.

In this talk, we explain what covariate adjustment is, how it works, when it may be useful to apply, and how to implement it. We will then discuss recent contributions to the field of covariate adjustment. In particular, we will touch on the role of data-adaptive methods (i.e., any data analysis method that adapts the data to learn structure). This includes sophisticated methods such as machine learning methods, with much flexibility to `adapt’ to the data, but also flexible parametric models with variable selection (e.g., stepwise variable selection, lasso, …), with or without the inclusion of splines.

Thursday, 07/Sept/2023 4:50pm - 5:10pm
ID: 127 / S40: 3
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Beyond the two-trials paradigm for generating pivotal evidence in drug development
Keywords: Pooling, replication, substantial evidence of effectiveness, two-trials convention

Combining clinical trials to generate pivotal evidence – case studies and reflections

Marc Vandemeulebroecke[1], Dieter Häring[1], Eva Hua[1], Xiaoling Wei[1], Dong Xi[2]
[1]Novartis; [2]Gilead; marc.vandemeulebroecke@novartis.com

In this talk, we present case studies that generated pivotal evidence using combined data from multiple pivotal trials in an overarching formal testing hierarchy. This goes beyond the traditional two-trials convention which requires independent pivotal evidence from (at least) two trials, separately. For each case study, we discuss the situation and rationale, the approach taken with its advantages and caveats, any experiences with health authorities, and the final outcomes such as resulting label claims.

Monday, 04/Sept/2023 4:10pm - 4:30pm
ID: 150 / S19: 1
Presentation Submissions - Regular Session (Default)
Topic of Submission: Clinical trial designs (adaptive and platform designs, external controls, …), Statistical modelling (regression modelling, prediction models, …), Personalized health care, Time-to-Event Analysis, Statistical hypothesis testing (covariate adjustment, nonparametric methods, multiple comparisons, …)
Keywords: Subgroup analysis, forest plots, shrinkage methods, bayesian shrinkage priors

Improved treatment effect estimation for time-to-event outcomes in subgroups displayed in forest plots based on shrinkage methods

Mar Vázquez Rabuñal[1,2], **Marcel Wolbers**[1], **Daniel Sabanés Bové**[1], **Kaspar Rufibach**[1]
[1]Hoffmann-La Roche Ltd, Basel, Switzerland; [2]ETH Zürich, Switzerland; mar.vazquez1998@gmail.com

In randomized controlled trials, the homogeneity of treatment effect estimates in pre-defined subgroups based on clinical, laboratory, genetic, or other baseline variables is frequently investigated using forest plots. Estimation of subgroup-specific treatment effects is typically based on the data from the respective subgroup only. However, the interpretation of these naive subgroup-specific treatment effect estimates requires great care because of the smaller sample size of the subgroups (implying large variability of the estimated effects) and the frequently large number of investigated subgroups. Bayesian analyses for treatment effect estimation have been discussed in the literature. These methods frequently focus on disjoint subgroups, whereas subgroups for different variables in forest plots are typically overlapping. We propose a general strategy for treatment effect estimation in subgroups for survival outcomes. We first build a flexible model based on all available observations including categorical covariates identifying the relevant subgroups and their interactions with the treatment group variable. Interaction terms are penalized using lasso or ridge regression to shrink subgroup-specific estimates towards the population treatment effect. Alternatively, we put a Bayesian shrinkage prior (the horseshoe prior) on the interaction terms. An advantage of the Bayesian approach is that it is straightforward to derive credible intervals for subgroup-specific estimates. In a second step, this model is marginalized to obtain treatment effect estimates (hazard ratios) for all subgroups. The non-collapsibility of the hazard ratio complicates marginalization and leads to marginalized survival curves for the treatment groups that are not proportional. To deal with these non-proportional survival curves we use the average hazard ratio corresponding to the odds-of-concordance to quantify the treatment effect.

The methods are illustrated with data from a large randomized clinical trial in follicular lymphoma and compared in an extensive simulation study. For all simulation scenarios, the overall mean-squared error (MSE) of all methods is drastically improved compared to naive subgroup-specific treatment effect estimates. The method based on the horseshoe prior performs slightly better in terms of bias, MSE and frequentist coverage of 95% credible intervals, compared to the other methods, in scenarios where only one of the subgrouping variables is associated with treatment effect heterogeneity. We are currently implementing all these methods in an R package which we plan to upload to CRAN.

Monday, 04/Sept/2023 4:10pm - 4:30pm
ID: 435 / S21: 1
Presentation Submissions - Regular Session (Default)
Topic of Submission: Free Contributions
Keywords: biostatistics, physicians, dentists, attitude, teaching

Biostatistics/Biometrics for physicians – essential or unnecessary? How do practicing physicians and dentists evaluate biostatistics? A cross-sectional survey

Maren Vens[1], **Nina Alida Hartmann**[2], **Inke Regina König**[1]
[1]Institut für Medizinische Biometrie und Statistik, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Universität zu Lübeck, Lübeck, Deutschland; [2]Institut für Sozialmedizin und Epidemiologie, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Universität zu Lübeck, Lübeck, Deutschland; m.vens@uni-luebeck.de
Note: A companion presentation to this contribution will be given at the "68. GMDS-JAHRESTAGUNG 2023".

Introduction

The aim of this project was to explore how physicians and dentists in Germany evaluate biostatistics in general and their education in this subject. Furthermore, the importance of the subject for the professional practice of physicians and dentists was determined and insights for teaching during and after medical school were gained.

Method

A total of 2700 physicians and dentists from Schleswig-Holstein were invited by mail to participate in an online survey and to provide sociodemographic data and information on their perception towards the subject of biostatistics in general, in relation to work and to teaching . Data were analyzed with descriptive methods.

Results

Response rate was 13.67%. 50.14% of participants received biostatistical training in medical school. 43.40% and 38.79% of participants reported that biostatistics was useful at study time and for their later career, respectively. Biostatistics was rated as difficult by 58.76%.

93.48% stated biostatistics being a necessary skill for a clinician involved in research, and 93.79% rated it as important for evidence-based medicine. 81.07% agreed that evidence-based medicine is important for clinical practice. Biostatistics was indicated as useful in their own work for evaluating marketing materials from pharmaceutical industry (88.64%), interpreting screening tests (87.88%), reading research publications for general professional interest (85.17%), using research publications to consider new therapeutic options (86.09%), in analyzing data from one's own research (91.89%), but only for 30.18% in clinical contact with patients.

20.00% rated the teaching in biostatistics from their own studies as still useful today. 65.22% would like to understand more about the subject while 86.96% received no further training after graduation. 53.04% of participants said they could do better if they understood more about biostatistics.

Discussion

Our study indicates that the majority of human and dental physicians in Schleswig-Holstein consider biostatistics to be difficult. However, they recognize the value of the subject for evidence-based medicine and research. More than 90%, or nearly one-third, expressed that biostatistics was a necessary skill for clinician scientists or practicing physicians, respectively. Biostatistics was indicated as helpful for a surprisingly large number of physician tasks, so that 13% got training in biostatistics even after graduation.

A large proportion of participants expressed dissatisfaction with biostatistics teaching from undergraduate days. Only one-fifth rated biostatistics teaching from those days as still useful today. Many reported uncertainty about their own biostatistics skills.

Conclusion

The results of this survey show that biostatistics is considered relevant in many areas of physician practice, but that many of the practitioners do not consider themselves well prepared and would like to understand more about biostatistics. Therefore, it seems reasonable to develop further training courses. These should focus precisely on the biostatistical concepts relevant to the medical/dental activities mentioned in the phase of early practical work. We therefore suggest to link the content within the framework of evidence-based medicine with clinically relevant medical topics, thus also increasing the motivation to learn biostatistics.

Thursday, 07/Sept/2023 12:00pm - 12:20pm
ID: 182 / **S68: 5**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Design of preclinical experiments
Keywords: Generalized Pairwise Comparisons, Permutation, Bootstrap, Preclinical trials

Potential of Generalized Pairwise Comparisons in pre-clinical studies

Johan Verbeeck
UHasselt, Belgium; johan.verbeeck@uhasselt.be

Pre-clinical studies are characterized by the measurement of multiple outcomes in small samples. To analyze these outcomes often rank-based or resampling permutation or bootstrap non-parametric test are applied. Recently, the non-parametric Generalized Pairwise Comparisons (GPC) methodology has attracted a lot of attention in clinical trials for its ability to evaluate multiple outcomes. Moreover, the GPC permutation test enjoys good small-sample properties. Curiously, this permutation test does not require re-sampling, a feature that can be extended to GPC bootstrap-based inference. GPC extends the Mann-Whitney test to multiple outcomes and shows great flexibility for the design of a study. It allows for any number and type of outcomes, allows for prioritizing the outcomes by clinical severity, allows for matched designs, allows for adding a threshold of clinical relevance and accounts for the correlation between the outcomes. In this talk, the generalized pairwise comparison ideas and concepts for small sample trials are introduced and critically evaluated for their applicability in preclinical trials.

Tuesday, 05/Sept/2023 11:20am - 11:40am
ID: 163 / **S23: 2**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Time-to-Event Analysis
Keywords: Non-Markov, Multistate models, Wild bootstrap, Confidence bands

Non-Markov non-parametric estimation of complex multistate outcomes after hematopoietic stem cell transplantation

Judith Vilsmeier[1], Sandra Schmeller[1], Daniel Fürst[2,3], Jan Beyersmann[1]
[1]Institute of Statistics, Ulm University, Ulm, Germany; [2]Institute of Clinical Transfusion Medicine and Immunogenetics Ulm, German Red Cross Blood Transfusion Service, Baden Wuerttemberg – Hessen and University Clinic Ulm; [3]Institute of Transfusion Medicine, Ulm University; judith.vilsmeier@uni-ulm.de

In many studies, probabilities of non-standard endpoints are of interest which are more complex than overall survival. One such probability is chronic GvHD- and relapse-free survival, the probability of being alive after stem cell transplantation, not suffering from chronic graft-versus-host disease (GvHD) and not having had a relapse with chronic GvHD being a recurrent event. Because the probabilities for such non-standard endpoints with recurrent events may not fall monotonically but may also rise again, one should not use a simple Kaplan-Meier estimator for the estimation of these probabilities, but the Aalen-Johansen estimator. One concern with this estimator is its consistency when the Markov assumption is not fulfilled, but Nießl et al. (2021) showed that the Aalen-Johansen estimator is in fact consistent even in non-Markov scenarios, as long as state occupation probabilities are estimated and the censoring is random. In some multistate models, it is also possible to estimate probabilities for complex, non-standard endpoints using linear combinations of Kaplan-Meier estimators. For these linear Kaplan-Meier combinations, we propose a wild bootstrap procedure for inference and to obtain confidence bands with the aim to compare the results with confidence bands obtained using the wild bootstrap technique for the Aalen-Johansen estimator in non-Markov scenarios (Bluhmki et al., 2018). In the proposed wild bootstrap procedure for the linear combinations of Kaplan-Meier estimators, the limiting distribution of the Nelson-Aalen estimator is approximated using the wild bootstrap and transformed via the functional delta method. This approach gives the same results as those of Liu et al. (2008) and Zhang et al. (2022) and is easily adaptable to different models. An advantage of the wild bootstrap is that the censoring may also be event-dependent, but then the Markov assumption must be fulfilled in the case of the Aalen-Johansen estimator. Using real data, confidence bands are generated using the wild bootstrap approach for the chronic GvHD- and relapse-free survival, since they provide an easy interpretive approach for e.g. two group comparisons. In addition, the coverage probabilities of confidence intervals and confidence bands generated by Efron's bootstrap and by the wild bootstrap are examined with simulations.

References:

1. A. Nießl, A. Allignol, J. Beyersmann and C. Mueller (2021): Statistical inference for state occupation and transition probabilities in non-Markov multi-state models subject to both random left-truncation and right-censoring. Econometrics and Statistics
2. T. Bluhmki, C. Schmoor, D. Dobler, M. Pauly, J. Finke, M. Schumacher and J. Beyersmann (2018): A wild bootstrap approach for the Aalen–Johansen estimator. Biometrics, 74: 977-985.
3. L. Liu, B. Logan and J.P. Klein (2008): Inference for current leukemia free survival. Lifetime data analysis, 14(4): 432–446.
4. X. Zhang, S.R. Solomon and C. Sizemore (2022): Inferences for current chronic graft-versus-host-disease free and relapse free survival. BMC Medical Research Methodology, 22: 318.

Monday, 04/Sept/2023 11:40am - 12:00pm
ID: 351 / **S6: 3**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), High dimensional data, genetic and x-omics data
Keywords: Gaussian copula graphical models, random graph models, Bayesian structural learning

Random graphical model of microbiome interactions in related environments

Veronica Vinciotti[1], **Ernst Wit**[2], **Francisco Richter**[2]
[1]University of Trento, Italy; [2]Università Svizzera italiana, Switzerland; veronica.vinciotti@unitn.it

Multivariate data are typically collected under different environments, such as different biological conditions or time points. The interest is often to discover the dependencies between the variables that are specific to each environment as well as structural similarities between the environments. We propose a computational approach for the joint inference of graphical models from different environments. A random graph generative model is introduced to capture relatedness at the structural level across the different environments. In addition, the model allows for the inclusion of external covariates at both the node and interaction levels, further adapting to the richness and complexity of high dimensional data from many application areas. We consider closely the inference of microbiota systems from metagenomic data for a number of body sites.

Tuesday, 05/Sept/2023 2:00pm - 2:20pm
ID: 131 / S30: 1
Presentation Submissions - Regular Session (Default)
Topic of Submission: Time-to-Event Analysis
Keywords: Additive hazard, Parametric modeling, Survival analysis, Time-to-event model.

A parametric additive hazard model for time-to-event analysis

Dina Voeltz[1,2], Amelie Forkel[2], Anke Schwandt[3], Oliver Kuss[4], Annika Hoyer[1]
[1]Bielefeld University, Germany, Germany; [2]Department of Statistics, Ludwig-Maximilians-University Munich, Germany; [3]Nuvisan GmbH, Neu-Ulm, Germany; [4]Institute for Biometrics and Epidemiology, German Diabetes Center, Leibniz Center for Diabetes Research at Heinrich Heine University Düsseldorf, Germany; dina.voeltz@uni-bielefeld.de

Regression models for time-to-event outcomes are predominantly fitted by the well-known Cox proportional hazard model. This is somewhat astonishing because its generic effect measure, the hazard ratio, has been repeatedly criticized in recent years. Points of concern were for instance the misleading interpretation as relative risk and its non-collapsibility.

A hazard-based model that overcomes most of these issues additive hazard model introduced by Aalen [1]. However, this approach [1] is rarely used in applied research, because it assumes a semi-parametric additive hazard as well as time-dependent covariates. Of course, these properties provide large flexibility in modelling, but also complicate parameter estimation considerably. As a partial remedy, Lin and Ying [2] proposed an additive hazard model for time-fixed covariates, but still insist on a semi-parametric additive hazard. Consequently, this renders computation and interpretation complicated, e.g., effect estimates of the Aalen approach are necessarily time-dependent and can only be given via graphs.

To overcome these limitations, we propose a parametric additive hazard model. This results in a number of advantages concerning interpretation, flexibility, possible model extensions, and technical implementation. For instance, being an essentially parametric model, it has survival, hazard and density functions directly available. Parameter estimation is straightforward and can be solved with any software that allows maximizing a user-written likelihood function.

We illustrate the model for different parametric distributional assumptions using data from the HALLUCA study [3] and show that the resulting parameter estimates and survival curves fit well with routinely used Kaplan-Meier curves. Further, results from a simulation study supports the finding that the approach works well in practice.

References

[1] Aalen OO. A linear regression model for the analysis of life times. Stat Med 1989;8:907-925.

[2] Lin DY, Ying Z. Semiparametric analysis of the additive risk model. Biometrika 1994;81:61-71.

[3] Bollmann A, Blankenburg T, Haerting J, Kuss O, Schütte W, Dunst J, Neef H. HALLUCA study. Survival of patients in clinical stages I-IIIb of non-small-cell lung cancer treated with radiation therapy alone. Strahlenther Onkol. 2004;180(8):488-96.

Wednesday, 06/Sept/2023 8:50am - 9:10am
ID: 262 / **S43: 2**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Estimands and causal inference
Keywords: Causal Inference, Clinical Trials, Simulations, Methods comparison

Outcomes truncated by death: a simulation study on the survivor average causal effect

Stefanie von Felten[1], Chiara Vanetta[1,2], Leonhard Held[1]
[1]University of Zurich, Switzerland; [2]MSD, Zurich, Switzerland; stefanie.vonfelten@uzh.ch

Continuous outcome measurements truncated by death present a challenge in RCTs, especially if mortality differs between groups. One way to deal with such situations is to estimate the survivor average causal effect (SACE). However, the SACE cannot be identified without non-testable assumptions.

We are involved in an ongoing RCT (EpoRepair) to evaluate the effect of high-dose recombinant human erythropoietin (Epo) on neurocognitive outcomes of very preterm infants with intraventricular hemorrhage. The primary outcome is IQ at 5 years of age. However, 15.7% of these vulnerable infants died until term equivalent age. Motivated by this trial we designed a simulation study to compare the SACE (using the estimator proposed by Hayden et al. 2005) with complete case analysis (of survivors) and multiple imputation of the missing outcomes (potentially less biased than complete case analysis). The outcome for the simulation study is mental development at two years of age (a secondary outcome in EpoRepair, already gathered), measured using the Bayley Scales of Infant and Toddler Development (BSID-III, cognition subscore). We chose 9 scenarios combining positive, negative and no treatment effect on the outcome (mean difference of 5, -5 or 0) and on survival (odds ratio of 2, 0.5 and 1). For each scenario we simulated 1300 data sets with 500 patients each, containing 4 baseline covariates (gestational age, head circumference at birth, 5 min Apgar score, socioeconomic score), the randomly allocated treatment (Epo vs. Placebo), the outcome and survival at 2 years of age (simulated observable data). Data simulation was based on summary statistics of baseline covariates, the correlation structure between them, and regression coefficients of the covariates on outcome/survival, taken from EpoRepair. In addition, we simulated outcome and survival counterfactuals under the corresponding other treatment for each patient. Each data set was analyzed by all three methods, and the treatment effect estimates based on the observable data were compared in terms of bias, mean square error (MSE) and coverage with regard to two types of true effects (estimands): (θ_1) the treatment effect on the outcome used in the simulation and (θ_2) the SACE derived from observed and counterfactual data.

In scenarios without a treatment effect on survival, all methods estimated similar treatment effects on the outcome which were close to θ_1 (and θ_2). In scenarios with a positive (negative) treatment effect on survival, complete case analysis estimated smaller (larger) positive and larger (smaller) negative treatment effects than the other methods, and small negative (positive) effects in case of no causal treatment effect on the outcome. This resulted in a bias for complete case analysis of around 10% with respect to θ_1 (and θ_2), the largest MSE and the lowest coverage. Although conceptually very different, estimates and performance measures were similar for the SACE estimator and multiple imputation.

There is currently limited awareness of the fact that outcomes truncated by death are not missing data in the usual sense. With our work we hope to promote awareness of this problem and to provide methodological knowledge of how it could be dealt with.

Tuesday, 05/Sept/2023 5:30pm - 5:50pm
ID: 407 / S41: 5
Presentation Submissions - Regular Session (Default)
Topic of Submission: Clinical trial designs (adaptive and platform designs, external controls, …), Statistical modelling (regression modelling, prediction models, …), Meta-Analysis and Systematic Reviews, Free Contributions
Keywords: Pediatric Extrapolation; Pediatric Development; Prior elicitation

Partial extrapolation in pediatric drug development using robust meta-analytic predictive priors, tipping point analysis and expert elicitation

Florian Voß[1], Morten Dreher[2], Elvira Erhardt[2], Heiko Müller[1], Oliver Sailer[2], Christian Stock[1]
[1]Boehringer Ingelheim Pharma GmbH & Co. KG, Ingelheim, Germany; [2]Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach, Germany; florian.voss@boehringer-ingelheim.com

Drug development in pediatric patients is often challenging. Pediatric trials have typically small sample sizes due to ethical considerations and feasibility reasons, particularly for rare diseases. As a result, they are often not powered to evaluate efficacy, but focus on safety aspects and pharmacology, and it is therefore difficult to reach conclusions on efficacy solely based on the pediatric trial. On the other hand, clinical development in pediatrics is initiated after clinical trials in adults have shown a positive benefit-risk profile. This allows to extrapolate and make use of the results obtained in adults to strengthen the evidence in the pediatric population (Gamalo et al, 2022). However, it is important to pre-specify how the external data from adults is used. A transparent and scientifically rigorous way to define the weight given to the external data and to assess the sensitivity of the results to the chosen weight is desirable. We show how Bayesian dynamic borrowing with robust meta-analytic predictive (MAP) priors, visualizations of tipping point analysis and expert elicitation can achieve this goal.

Our approach is motivated and illustrated by a real example in a rare pediatric disease with a continuous endpoint and several available trials in related adult indications. During the planning stage, formal prior elicitation is conducted with a panel of clinical experts. Results of a meta-analysis in adult patients are shown to them in a first step to assess the robustness across adult trials. A tipping point analysis, as used by Best at al. (2021) with slight modification, is used to visualize for a wide range of hypothetical result of the pediatric trial and different one-sided evidence levels how much weight is needed for the informative component of the robust MAP prior to infer that the treatment is efficacious. Next, the panel is asked about conclusions they would make from the available evidence in different scenarios and the weights they would put on the evidence from the trials in adults considering their believe in the applicability of the adult data, tipping point analyses and other operating characteristics. The pre-specified weight is then derived via a roulette method (Gosling 2018) based on the experts' opinions and used for the primary analysis while the tipping point analysis serves as a sensitivity analysis for the impact of the chosen weight on the conclusion.

We also discuss our approach in the context of the new draft guidance on pediatric extrapolation (ICH, 2022) and introduce a publicly available R package called "tipmap" which facilitates the implementation of the described approach.

References:

1. Best N, et al. Assessing efficacy in important subgroups in confirmatory trials: An example using Bayesian dynamic borrowing. Pharm Stat. 2021;20(3):551-562.
2. Gamalo M, et al. Extrapolation as a Default Strategy in Pediatric Drug Development. Ther Innov Regul Sci. 2022;56:883–894.
3. Gosling 2018. SHELF: The Sheffield Elicitation Framework. In: Dias, L., Morton, A., Quigley, J. (eds) Elicitation. International Series in Operations Research & Management Science, vol 261. Springer, Cham. https://doi.org/10.1007/978-3-319-65052-4_4
4. ICH guideline E11A on pediatric extrapolation. Draft version URL: https://www.ema.europa.eu/en/documents/scientific-guideline/draft-ich-guideline-e11a-pediatric-extrapolation-step-2b_en.pdf

Tuesday, 05/Sept/2023 11:00am - 11:20am
ID: 448 / **S25: 1**
Presentation Submissions - Featured Session
Featured Sessions: Young Statisticians Sessions and Panel Discussion
Keywords: cancer screening programmes, lead time, length time bias, survival analysis

Evaluating cancer screening programmes using survival analysis

Bor Vratanar, Maja Pohar Perme
Faculty of Medicine, Slovenia; bor.vratanar@gmail.com

Cancer screening is a programme for medical screening of asymptomatic people who are at risk of developing cancer. Typically, participants are regularly screened every few years using blood tests, urine tests, medical imaging, or other methods. Among cases who are screened regularly some are diagnosed with cancer based on screening tests (screen-detected cases) and some based on symptoms appearing in the interval between two consecutive screening tests (interval cases). The hypothesis is that the screening programmes improve chances of survival for screen-detected cases as these cases are diagnosed and treated at an earlier stage of the disease compared to counterfactual scenario where their cancer would have been detected based on symptoms. We would like to test this hypothesis empirically. So far, the problem has been tackled by comparing the survival functions of screen-detected cases and interval cases. Realizing that the direct comparison between these two groups would result in biased results, previous research focused on parametric solutions to remove the bias. We argue that the problem lies elsewhere – that this comparison, in fact, does not reflect the question of interest. Therefore, in this study, we precisely define the contrast corresponding to the hypothesis defined above. Since the contrast of interest refers to hypothetical quantities, we discuss which data and under what assumptions can be used for estimation. We also propose a non-parametric framework for evaluating the effectiveness of cancer screening programmes under certain assumptions. The proposed ideas are illustrated using simulated data. The problem is motivated by the need to evaluate breast cancer screening programme in Slovenia.

Wednesday, 06/Sept/2023 11:40am - 12:00pm
ID: 323 / S50: 3
Presentation Submissions - Invited Session
Invited Sessions: Covariate Adjustment in RCTs: Translating theory into practice within a pharmaceutical company via a data challenge
Keywords: Covariate Adjustment

Participating in a data challenge on covariate adjustment in RCTs

Craig Wang
Novartis, Switzerland; craig.wang@novartis.com

Covariate adjustment in randomized trials is a currently a topic of interest for health authorities, highlighted by a recent FDA guidance document and EMA qualification opinion on a particular type of adjustment.

An internal “Covariate Adjustment Challenge” was run within the Analytics department at Novartis. 23 participating teams were given access to data from five prior studies in a particular indication. The aim was to propose either a single "super" covariate or a pre-specified set of baseline covariates that could be used in covariate-adjusted analyses of key endpoints in a subsequent study in the same indication. Upon the “test data” becoming available, teams were scored according to the gain in precision from their proposed adjustment compared to an unadjusted analysis.

This talk will discuss the challenge from a participant’s perspective, including the overall strategy taken, the specific methods used, implementation, as well as how the proposal could be communicated to clinical colleagues.

Tuesday, 05/Sept/2023 5:10pm - 5:30pm
ID: 298 / **S41: 4**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Real world data and evidence
Keywords: Bayesian dynamic borrowing; Augment control arm, Bayesian bootstrap, Minimize mean squared errors

Dynamic borrowing to minimize mean squared error and inference with Bayesian bootstrap

Jixian Wang[1], **Ram Tiwari**[2]
[1]BMS, Switzerland; [2]BMS, US; jixian.wang@bms.com

Dynamic borrowing to augment the control arm of a small RCT leveraging external data sources has been increasingly used. The key step in dynamic borrowing is determining the amount of borrowing based on similarity of controls from the trial and the external data sources. A simple approach to this task is the empirical Bayesian approach which maximizes the marginal likelihood (maxML) of the amount of borrowing. We propose an alternative that determines the amount of borrowing to minimize the mean squared error (minMSE) of the estimator combining both sources. We derive a simple algorithm using Bayesian bootstrap for determining the amount of borrowing that minimizes the posterior MSE based on posterior means and variances. It can also be used with a pre-adjustment on the external controls for population difference between the two sources using, e.g., inverse probability weighting. We show that the minMSE rule has similar asymptotic properties as the maxML rule, which leads to either full or no borrowing, depending on weather the (adjusted) means of control outcome from the two sources are the same or not. Statistical inference can be made using the Bayesian bootstrapped posterior sample, or approximate asymptotic normal distribution. A simulation study is performed to compare the min MSE rule with the maxML rule. Generally, the minMSE rule leads to smaller MSE than that from the maxML rule, but its coverage of frequentist 95 percent confidence interval may be better or worse than that of the maxML rule, depending on multiple factors . Our approach is very easy to implement and computationally efficient. The approach is illustrated by an example of borrowing controls of an Acute Myeloid Leukemia trial from another study.

Monday, 04/Sept/2023 11:20am - 11:40am
ID: 149 / S5: 2
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Finding the right dose – Project Optimus and beyond

Towards efficient dose-escalation guidance of multi-cycle cancer therapies

Sebastian Dragos Weber[1], Lukas Andreas Widmer[1], Yunnan Xu[2], Hans-Jochen Weber[1]
[1]Novartis Pharma AG, Switzerland; [2]Novartis Pharma AG, USA; sebastian.weber@novartis.com

Treatment of cancer has rapidly evolved over time in quite dramatic ways, for example from chemotherapies, targeted therapies to immunotherapies and chimeric antigen receptor T-cells. Nonetheless, the basic design of early phase I trials in oncology still follows pre-dominantly a dose-escalation design monitoring the safety of the first treatment cycle only. With toxicities occurring at later stages beyond the first cycle and the need to treat patients over multiple cycles, the focus on the first treatment cycle only is becoming a limitation in nowadays multi-cycle treatment therapies. Here we introduce a multi-cycle time-to-event model allowing guidance of dose-escalation trials studying multi-cycle therapies. The challenge lies in balancing the need to monitor safety of longer treatment periods with the need to continuously enroll patients safely. We introduce in this work a multi-cycle time to event model which is formulated as an extension to existing approaches like the escalation with overdose control principle. The model is motivated from a drug development project and evaluated in a simulation study.

Tuesday, 05/Sept/2023 12:00pm - 12:20pm
ID: 239 / **S27: 4**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Statistical issues in health care provider comparisons
Keywords: patient reported outcomes, quality indicators, quality assurance

Patient surveys for assessing medical treatment quality

Felix Weidemann
IQTIG, Germany; felix.weidemann@iqtig.org

The IQTIG (Institute for Quality Assurance and Transparency in Healthcare) develops and evaluates quality indicators in order to compare treatment performance of stationary and ambulant healthcare providers in Germany. Since 2022, these indicators use survey responses from patients as an additional data source. This allows to measure treatment aspects which are part of high quality care from a patient perspective (patient reported outcomes). This includes, for instance, pain therapy or patient information.

Quality indicators for patient reported outcomes are formed as index measures, which aggregate several quality aspects and associated survey items. To evaluate such index measures, the IQTIG developed a statistical procedure with the aim to provide a fair and robust quantitative assessment of healthcare providers. The statistical procedure as well as its quantitative results should be easily comprehendible and informative for both providers and patients.

The statistical procedure is based on a hierarchical formative-reflective quality indicator model. All underlying quality attributes are estimated through a Bayesian beta-binomial model. The statistical procedure was examined by simulation studies focussing on the classification properties, i.e. the ability to detect providers with quality deficits.

In 2023 the IQTIG computed first results for the developed patient reported outcome indicators. Within a trial phase, these results are reported back to providers to provide feedback and allow comparison.

Tuesday, 05/Sept/2023 2:00pm - 2:20pm
ID: 188 / **S31: 1**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Interpretable machine learning in biostatistics: Methods, applications and perspectives

Interaction difference test for prediction models

Thomas Welchowski
University Hospital Bonn, Germany; welchow@imbie.uni-bonn.de

Machine learning research focuses on the improvement of prediction performance. During the past decade, major advances were made in the field of deep learning with imaging, audio or video data and ensemble models like bagging or boosting with matrix type data. These so called black-box models flexibly adapt to the given data and involve fewer assumptions about the data generating process than standard methods like linear regression or single decision trees. However, due to their increased complexity, black-box models are more difficult to interpret. To address this issue, techniques for interpretable machine learning have been developed; yet there is still a lack of methods to reliably identify interaction effects between predictors under uncertainty.

In this work we present a model-agnostic asymptotic hypothesis test for the identification of interaction effects in black-box machine learning models. The null hypothesis assumes that a given set of covariates does not contribute to interaction effects in the prediction model. The test statistic is based on the difference of variances of partial dependence functions with respect to the original black-box predictions and the (more restrictive) predictions under the null hypothesis. Properties of the proposed test statistic (in particular its power) were explored in simulations of linear and nonlinear models.

The proposed hypothesis test can be applied to any black-box prediction model under suitable consistency assumptions, and the null hypothesis of the test can be flexibly specified/modified according to the research question of interest. Furthermore, the test is computationally fast to apply as the null distribution does not require resampling and/or re-fitting black-box prediction models.

Wednesday, 06/Sept/2023 11:40am - 12:00pm
ID: 471 / **S51: 4**
Presentation Submissions - Featured Session
Featured Sessions: IBS-DR/ROeS Award session
Keywords: meta-regression, multivariate analysis, cluster-robust estimators, mixed-effects models

Cluster-robust estimators for multivariate mixed-effects meta-regression

Thilo Welz[1,2], **Wolfgang Viechtbauer**[3], **Markus Pauly**[1,4]
[1]TU Dortmund University; [2]Daiichi Sankyo Europe; [3]Maastricht University; [4]UA Ruhr, Research Center Trustworthy Data Science and Security; welz.thilo@gmail.com

Meta-analyses frequently include trials that report multiple outcomes based on a common set of study participants. These outcomes will generally be correlated. Cluster-robust variance-covariance estimators are a fruitful approach for synthesizing dependent outcomes. However, when the number of studies is small, state-of-the-art robust estimators can yield inflated Type 1 errors. Therefore, two new cluster-robust estimators are presented, in order to improve small sample performance. For both new estimators the idea is to transform the estimated variances of the residuals using only the diagonal entries of the hat matrix. The proposals are asymptotically equivalent to previously suggested cluster-robust estimators such as the bias reduced linearization approach. The methods are applied to a dataset of 81 trials examining overall and disease-free survival in neuroblastoma patients with amplified versus normal MYC-N genes. Furthermore, their performance is compared and contrasted in an extensive simulation study. The focus is on bivariate meta-regression, although the approaches can be applied more generally.

Tuesday, 05/Sept/2023 12:00pm - 12:20pm
ID: 424 / S28: 4
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), Meta-Analysis and Systematic Reviews
Keywords: network meta-analysis, non-proportional hazards

Nonproportional Hazards in Network Meta-Analysis: Efficient Strategies for Model Building and Analysis

Anna Wiksten[1], **Tiina Kirsilä**[2], **Hans-Peter Piepho**[3], **Zehua Zhou**[1]
[1]Bristol Myers Squibb, Switzerland; [2]Novartis Pharma AG, Switzerland; [3]University of Hohenheim, Germany;
anna.wiksten@bms.com

Cancer therapies with different mode of actions often show different time of efficacy onset. This leads to survival data violating the proportional hazards assumption. Network meta-analysis (NMA) of such data should acknowledge this. Two suitable approaches are fractional polynomial (FP) models and piece-wise constant (PWC) models. These types of models can be difficult to fit in practice, and there is a need for efficient model building and selection strategies.

In this presentation we will present a structured model building strategy for network meta-analysis of digitized Kaplan-Meier curves. We will show how the initial model selection process can be done using frequentist modelling with arm-based NMA parameterization and then selected model and the uncertainty is evaluated using Bayesian methods as presented in Jansen (2011). The proposed model building strategy was published in Wiksten et al (2020).

We will also present technical solutions how new studies can be easily added to existing network using R markdown template.

Wednesday, 06/Sept/2023 9:10am - 9:30am
ID: 345 / S48: 3
Presentation Submissions - Regular Session (Default)
Topic of Submission: Clinical trial designs (adaptive and platform designs, external controls, …)
Keywords: Bayesian design, Cluster randomised trials, Assurance, Sample size determination, Hybrid approach

Sample size calculations for cluster randomised trials using assurance

Sarah Faye Williamson[1], Svetlana V. Tishkovskaya[2], Kevin J. Wilson[3]
[1]Biostatistics Research Group, Newcastle University, UK; [2]Faculty of Health and Care, Lancashire Clinical Trials Unit, University of Central Lancashire, UK; [3]School of Mathematics, Statistics & Physics, Newcastle University, UK;
faye.williamson@newcastle.ac.uk

Sample size determination for cluster randomised trials (CRTs) is challenging because it requires robust estimation of the intra-cluster correlation coefficient (ICC). Typically, the sample size is chosen to provide a certain level of power to reject the null hypothesis in a two-sample hypothesis test. This relies on the minimal clinically important difference (MCID) and estimates of the overall standard deviation, the ICC and, if cluster sizes are assumed to be unequal, the coefficient of variation of the cluster size. Varying any of these parameters can have a strong effect on the required sample size. The ICC can be particularly challenging to estimate and, if the value used in the power calculation is far from the unknown true value, can lead to trials which are substantially over- or under-powered.

In this talk, we present a hybrid approach which uses the Bayesian concept of assurance (or expected power) to determine the sample size for a CRT in combination with a frequentist analysis. Assurance is a robust alternative to traditional power which incorporates the uncertainty on key parameters through prior distributions. We suggest specifying prior distributions for the overall standard deviation, ICC and coefficient of variation of the cluster size, while still utilising the MCID.

This approach is motivated by a parallel-group CRT in post-stroke incontinence. Although a pilot study was conducted for this trial, the resulting ICC estimate was of low precision and could not be used as a reliable source for the sample size calculation. We illustrate the effects of redesigning this trial using the hybrid approach and compare the results to those obtained from a standard power calculation. The impacts of misspecifying the ICC prior distribution are also considered.

Monday, 04/Sept/2023 5:30pm - 5:50pm
ID: 269 / **S21: 5**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Statistical modelling (regression modelling, prediction models, …), Real world data and evidence, Time-to-Event Analysis
Keywords: alien species invasions, relational event models, counting processes, first records data, ecology

Studying global alien species invasions between 1880 and 2005 with relational event models

Ernst C. Wit, Ruta Juozaitiene, Martina Boschi
Universita della Svizzera italiana, Switzerland; wite@usi.ch

Spatio-temporal interactive processes, such as alien species invasions, play a key role in ecology. Existing methods studying such processes often simplify the dynamic structure or the complex interactions of the ecological drivers. In this talk we show how to use relational event modelling (REM) for analysing patterns of ecological interaction processes at large spatial scales including time-varying variables that drive these dynamics. REM relies on temporal interaction dynamics, that encode sequences of relational events connecting a sender node to a recipient node at a specific point in time. We apply REM to the spread of alien species around the globe between 1880 and 2005, following accidental or deliberate introductions into geographical regions outside of their native range. In this context, a relational event represents the new occurrence of an alien species given its former distribution. The application of relational event models to the first reported invasions of 4835 established alien species outside of their native ranges from four major taxonomic groups enables us to unravel the main drivers of the dynamics of the spread of invasive alien species. Combining the alien species' first records data with other spatio-temporal information enables us to discover which factors have been responsible for the spread of species across the globe. Besides the usual drivers of species invasions, such as trade, land use, and climatic conditions, we also find evidence for species-interconnectedness in alien species spread. Relational event models offer the capacity to account for the temporal sequences of ecological events such as biological invasions and to investigate how relationships between these events and potential drivers change over time.

Thursday, 07/Sept/2023 11:20am - 11:40am
ID: 158 / **S70: 3**
Presentation Submissions - Regular Session (Default)
Topic of Submission: COVID pandemic, Epidemiology
Keywords: COVID-19, hospitalization incidence, nowcasting, methods comparison, forecast combination

Collaborative nowcasting of COVID-19 hospitalization incidences in Germany

Daniel Wolffram[1,2], **Melanie Schienle**[1,2], **Johannes Bracher**[1,2]
[1]Karlsruhe Institute of Technology, Germany; [2]Heidelberg Institute for Theoretical Studies; daniel.wolffram@kit.edu

Real-time surveillance data are a crucial element in the response to infectious disease outbreaks. However, their interpretation is often hampered by delays in data collection and reporting, which bias recent values downward and can obscure current trends. Statistical nowcasting can be employed to correct these biases and enhance situational awareness. This talk summarizes a pre-registered real-time assessment of eight nowcasting approaches, applied by independent research teams to German 7-day hospitalization incidences. All methods were applied from 22 November 2021 to 29 April 2022, each day issuing probabilistic nowcasts for the current and 28 preceding days. Nowcasts were collected in a public repository and displayed in a dashboard. Moreover, mean and median ensembles were generated. All compared methods were able to remove a large part of the biases due to delays. Most teams underestimated the importance of long delays, though, resulting in nowcasts with a slight downward bias. Also, the uncertainty intervals of most methods were too narrow. Averaged across horizons, the best performance was achieved by a model using case incidences as a covariate and accounting for longer delays than the other approaches. For the most recent days, which are often considered particularly relevant, the mean ensemble performed best.

Wednesday, 06/Sept/2023 8:30am - 8:50am
ID: 286 / **S45: 1**
Presentation Submissions - Regular Session (Default)
Topic of Submission: High dimensional data, genetic and x-omics data
Keywords: gene set analysis, gene set enrichment, over-optimism, variability of results, questionable research practices

Over-optimism in gene set analysis: How does the choice of methods and parameters influence the detection of differentially enriched gene sets?

Milena Wünsch[1,3], Christina Nießl[1], Ludwig Christian Hinske[2], Anne-Laure Boulesteix[1]
[1]Institute for Medical Information Processing, Biometry and Epidemiology, University of Munich, Munich, Germany; [2]Institute for Digital Medicine, University Hospital of Augsburg, Augsburg, Germany; [3]Munich Center for Machine Learning (MCML), Munich, Germany; milena.wuensch@ibe.med.uni-muenchen.de

Gene set analysis, a popular approach for analyzing high-throughput gene expression data, aims to identify sets of related genes that show significantly enriched or depleted expression patterns between two opposed conditions. In addition to the multitude of methods available for this task, the user is typically left with many options when creating the required input and specifying the internal parameters of the chosen method. This flexibility might not only make it difficult to get a clear overview of all steps required to conduct gene set analysis but also entice users to produce the most preferable results in a “trial-and-error” manner. While this procedure seems natural at first glance, it can be viewed as a form of “cherry-picking” and cause an over-optimistic bias in the results. Since the method and its underlying parameters are exceedingly fitted to the given gene expression dataset, the results may not be replicable with a different dataset, leading to a loss in validity – a problem that has attracted a lot of attention in the context of classical hypothesis testing. In this talk, we aim to raise awareness of this type of over-optimism in the more complex context of gene set analysis. First, we give an overview of the general theoretical background of gene set analysis and summarize the methodology behind seven popular methods classified as Over-Representation Analysis or Functional Class Scoring. Second, we discuss the practical aspects of applying these methods, which are implemented either in popular R packages, such as clusterProfiler and GOSeq, or in web-based applications, such as GSEA. Finally, to address the problem of over-optimism, we mimic a hypothetical researcher engaging in the systematic selection of the underlying options for the purpose of optimizing the results. More precisely, we perform optimization for three metrics, each within two real gene expression datasets frequently used in benchmarking. In addition to optimizing these metrics for the true sample labels of the gene expression datasets, we repeat this procedure for ten randomly generated permutations of the sample labels. Our study suggests that for most gene set analysis methods, the set of options left to the user can lead to a particularly high variability in the number of differentially enriched gene sets as well as in the ranking of the gene sets in the corresponding results. This underlines the risk of selective reporting and over-optimistic results in the context of gene set analysis.

Tuesday, 05/Sept/2023 4:30pm - 4:50pm
ID: 128 / S40: 2
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Beyond the two-trials paradigm for generating pivotal evidence in drug development
Keywords: familywise error rate, combined analysis, submissionwise error rate, two-trials convention

When convention meets practicality: Combined analysis testing under the two-trials convention

Dong Xi[1], Frank Bretz[2], Willi Maurer[2]
[1]Gilead Sciences, United States of America; [2]Novartis AG, Basel, Switzerland; dong.xi1@gilead.com

Regulatory guidance suggests controlling the family-wise error rate (FWER) in confirmatory clinical trials. The two-trial paradigm represents a further requirement to demonstrate efficacy in a clinical submission: A statistically significant outcome in at least two adequate and well-controlled clinical trials. Within each trial, different endpoints may require different sample size to achieve the adequacy of power. Sometimes the sample size driven by one endpoint could be twice as large as that required by other endpoints. These unbalanced requirements of resources in a single trial are amplified under the two-trial convention and may lead to financial and logistical challenges for the trial sponsor. It is, therefore, often of interest to combine the data from the two trials for an endpoint to make a confirmatory claim without doubling the sample size. However, it remains unclear what approaches could be used to manage multiplicity adjustments for the combined analysis using data from two identically designed trials. In this talk, we provide principles of controlling the submission-wise error rate (SWER) with combined analyses when success claims for endpoints tested in individual trials should be based on significance in both trials. We also discuss examples of SWER under other requirements where success claims could be based on significance from a single trial.

References

1. Bretz, F., Maurer, W., & Xi, D. (2019). Replicability, reproducibility, and multiplicity in drug development. Chance, 32(4), 4-11.
2. Bretz, F., & Xi, D. (2019). Commentary on “Statistics at FDA: Reflections on the Past Six Years”. Statistics in Biopharmaceutical Research, 11(1), 20-25.

Tuesday, 05/Sept/2023 11:40am - 12:00pm
ID: 309 / **S23: 3**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Time-to-Event Analysis
Keywords: Multi-state models, semi-Markov processes, event history data, inter-current events, estimands

Analysis of Time to Treatment Responses: An Application of a Multi-State Model using Semi-Markov Process

Lillian Yau[1], **Meng Cao**[2]
[1]Novartis AG, Basel, Switzerland; [2]Novartis Pharmaceuticals Corporation, East Hanover, USA; lillian.yau@novartis.com

In clinical studies, multiple correlated survival times arise naturally when a patient can experience multiple events over the course of a study. Quantities of interests include probabilities that patients experience certain events, or amount of time elapsed before certain events occur. In the estimand framework, often one event is of interests, and many others are considered inter-current. Data analysis in such setups requires simplifying assumptions. For example, when the survival time of interest is a sum of multiple survival times, the intermediate events are usually ignored, and the total survival time is modeled directly. Moreover, when interests change to a different event, separate models must be fitted by reversing the roles of the event of interest and the inter-current events.

One solution to this is to model such data with multi-state models (MSM). MSMs support the estimand framework by defining all the relevant events—including all possible intercurrent events—as states in an MSM. Complete patient trajectories can be visualized through these states. MSMs also allow us to compare different estimand strategies for addressing the inter-current events by identifying corresponding states and their transitions. As a result, MSMs extend traditional univariate time-to-event analyses methods.

Many approaches to analyzing MSMs have been proposed. Markov processes have gained popularity for their algebraic simplicity. However, the Markovian assumption implies exponentially distributed survival times, which usually is too restrictive. Alternatively, non-parametric methods based on counting processes are a common choice; yet numerical trackability becomes an issue when the number of states increases.

In this presentation we introduce semi-Markov processes (SMPs), a versatile tool for MSMs. They generalize Markov processes by allowing survival times between states to have nearly any distributions—parametric or non-parametric.

The theory of SMPs was fully developed by the 1960s. One important mathematical concept that enables the derivations of the quantities of interests in MSMs is Laplace transform (LT). In particular, the close relationship between LTs and the characteristic functions of distributions allows complex computations from the time domain, for instance, convolutions of multiple integrals of distributions of survival times, to be translated to simpler algebraic manipulations in the frequency domain, in this case, products of LTs. The LTs of quantities of interests are then inverted back to the time domain analytically or numerically.

The numerical inversion of LTs had been an obstacle for the use of SMPs in the past. With the advances of computational powers, the SMPs have gained recognition, and have been utilized for analyzing MSMs for a variety of applications. In addition, a new R package “smproc” further alleviates the computational burden of implementing SMPs for data analysts.

As an illustration, we present an exploratory data analysis where we use an SMP to characterize different stages of treatment responses. The data were collected in a recently completed pivotal phase III randomized study, testing a new compound against a standard of care. The primary read-out of the study demonstrated the superiority of the new compound, which has since been approved by many health authorities including the FDA and EMA.

Thursday, 07/Sept/2023 8:30am - 9:10am
ID: 302 / S57: 1
Presentation Submissions - Invited Session
Invited Sessions: Causal inference and the art of asking meaningful questions
Keywords: causal inference, competing risks

Causal inference with competing events

Jessica Young
Harvard Medical School & Harvard Pilgrim Health Care Institute, United States of America; jyoung@hsph.harvard.edu

A competing (risk) event is any event that makes it impossible for the event of interest in a study to occur. For example, cardiovascular disease death is a competing event for prostate cancer death because an individual cannot die of prostate cancer once he has died of cardiovascular disease. Various statistical estimands have been posed in the classical competing risks literature, most prominently the cause-specific cumulative incidence, the marginal cumulative incidence, the cause-specific hazard, and the subdistribution hazard. Here we will discuss the interpretation of counterfactual contrasts in each of these estimands under different treatments and consider possible limitations in their interpretation when a causal treatment effect on the event of interest is the goal and treatment may affect future event processes. In turn, we argue that choosing a target causal effect in this setting fundamentally boils down to whether or not we choose to be satisfied estimating total effects, that capture all mechanisms by which treatment affects the event of interest, including via effects on competing events. When we deem the total effect insufficient to answer our underlying question, we consider alternative targets of inference that capture treatment mechanism for competing event settings, with emphasis on the recently proposed separable effects.

Thursday, 07/Sept/2023 10:40am - 11:00am
ID: 143 / **S66: 1**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Advanced survival analysis
Keywords: Accelerated life testing, tampered random variable model, simulation study, chronic diseases, type 2 diabetes

Modelling chronic disease mortality by methods from accelerated life testing

Marina Zamsheva[1], Andreas Wienke[1], Oliver Kuss[2,3]
[1]Institute of Medical Epidemiology, Biostatistics, and Informatics, Interdisciplinary Center for Health Sciences, Medical Faculty of the Martin-Luther-University Halle-Wittenberg, Germany; [2]German Diabetes Center, Leibniz Institute for Diabetes Research at Heinrich Heine University Düsseldorf, Institute for Biometrics and Epidemiology, Düsseldorf, Germany; [3]Centre for Health and Society, Medical Faculty and University Hospital Düsseldorf, Heinrich Heine University Düsseldorf, Germany;
Marina.Zamsheva@uk-halle.de

Methods of accelerated life testing (ALT) are widely used in reliability theory for estimating lifetime of technical devices. To this task, items are exposed systematically to higher stress levels of, e.g. temperature, voltage or pressure. This approach is more efficient by providing more failures in shorter observation time and information for higher stress levels is used to estimate lifetime under normal conditions.

We propose to use these ideas in the epidemiology of chronic diseases by conceptualizing the diagnosis of a chronic disease as exposing a person to a higher stress level, thus potentially shortening its residual lifetime, or, equivalently, accelerating its time to death. We use the tampered random variable model of Degroot and Goel (1979) and Gompertz distributions to model mortality from type 2 diabetes using data from the population-based CARLA cohort. The TRV model correctly accounts for the semi-competing risk structure in the data, allows entry into the cohort at higher ages, and uses information from prevalent as well as incident cases. In addition, using parametric distributions offers reporting model results on the original time, rather than on the hazard scale. In an extension of the model we also allow the age at diabetes diagnosis to be observed not exactly, but only in an interval. Model parameters can be estimated straightforwardly by maximum likelihood, and we give some preliminary results of a simulation study showing our approach working well.

Tuesday, 05/Sept/2023 4:10pm - 4:30pm
ID: 129 / **S40: 1**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Beyond the two-trials paradigm for generating pivotal evidence in drug development
Keywords: pooled analysis, power, regulatory guideline, type I error control, two-trials convention

Comparison of one-trial and two-trial paradigms in drug assessment

Stella Jinran Zhan[1], **Cornelia Ursula Kunz**[2], **Nigel Stallard**[1]
[1]University of Warwick, United Kingdom; [2]Boehringer Ingelheim Pharma GmbH & Co. KG, Germany;
stella.zhan@warwick.ac.uk

Regulators have usually required at least two significant independent pivotal trials as substantial evidence of the effectiveness of a new drug. This standard requirement is known as the two-trial paradigm and has remained the conventional approach for decades.

However, to adhere to this standard rule, sponsors commonly design and conduct two identical trials. As an alternative approach, one might combine the data from the two trials into a single trial (one-trial paradigm) and obtain a higher power. It can be shown that this method would ensure the same level of type I error protection as the two-trial paradigm only under a specific scenario, but there is little investigation on the type I error protection over the whole null region.

In this talk, we compare the two-trial paradigm to the one-trial paradigm to better understand the regulatory decision-making in the assessment of drugs' effectiveness, specifically what statistical errors the regulators are trying to protect against. We consider scenarios in which the two trials are conducted in identical or different populations as well as with equal or unequal size. With identical populations, the results show that a single trial provides better type I error protection and higher power. Conversely, with different populations, although the one-trial rule is more powerful in some cases, it does not always protect against the type I error.

Reference

Zhan, SJ, Kunz, CU, Stallard, N. Should the two-trial paradigm still be the gold standard in drug assessment? *Pharmaceutical Statistics*. 2023; 22(1): 96-111.

Thursday, 07/Sept/2023 12:00pm - 12:20pm
ID: 324 / S70: 5
Presentation Submissions - Regular Session (Default)
Topic of Submission: COVID pandemic, Statistical modelling (regression modelling, prediction models, …), Epidemiology, High dimensional data, genetic and x-omics data

Bayesian Poisson Regression and Tensor Train Decomposition Model for Learning Mortality Pattern Changes during COVID-19 Pandemic

Wei Zhang, Antonietta Mira, Ernst C Wit
Università della Svizzera italiana, Switzerland; wei.zhang@usi.ch

COVID-19 has led to excess deaths around the world, however it remains unclear how the mortality of other causes of death has changed during the pandemic. Aiming at understanding the wider impact of COVID-19 on other death causes, we use an Italian dataset that consists of monthly mortality counts of different causes of death staring from pre-COVID-19 era to June 2020. Due to the high dimensional nature of the data, we developed a model which combines the conventional Poisson regression with tensor train decomposition to explore the lower dimensional structure of the data. We take a Bayesian approach and impose priors on model parameters. The posterior inference is made using an efficient Metropolis-Hastings within Gibbs algorithm. Our method provides informative interpretations that conform to our hypothesis of the relationship between COVID-19 and other causes of death in addition to the Poisson regression.

Tuesday, 05/Sept/2023 2:00pm - 2:20pm
ID: 468 / **S32: 1**
Presentation Submissions - Featured Session
Featured Sessions: Young Statisticians Sessions and Panel Discussion
Keywords: exploratory analysis, variable selection, type-I error control, knockoff framework

Enhancing replicability of exploratory variable selections based on clinical trial data

Manuela Rebecca Zimmermann, Mark Baillie, Matthias Kormaksson, David Ohlssen, Konstantinos Sechidis
Novartis, Switzerland; manuela.zimmermann@novartis.com

Clinical trials are at the core of clinical development and a major driver of costs in the pharmaceutical industry. However, their primary purpose is to answer only narrowly defined scientific questions. While detailed prespecification of such questions is required for regulatory purposes, it seems wasteful not to make more use of clinical trial data, given the enormous efforts required to obtain such data sets and their generally high validity. Indeed, there is a great interest in re-using clinical trial data to support scientific discovery, e.g. to identify prognostic measures of disease or biomarkers that predict treatment efficacy. The issue with the vast majority of such exploratory analyses, however, is that they can be overly optimistic about positive findings. Indeed, most exploratory analyses do not account for multiplicity, and thus contribute to the current replicability crisis. In the context of clinical development, this lack of control for false discoveries (type-I errors) results in increased patient burden, unnecessary research efforts, and avoidable costs.

The academic literature offers a robust and versatile framework for exploratory variable selection analyses under strict type-I error control, the *knockoff* framework. This framework can handle high dimensional data in a model-agnostic manner, which makes it a prime candidate for exploratory analyses in clinical development settings. However, its operation characteristics in practically relevant settings are largely unknown and generally depend on the myriad choices that can be made when applying the framework. We raise awareness for practical considerations, exemplify common pitfalls, and demonstrate practical performance of the knockoff framework in real case studies of Phase III trials. In addition to the methodology transfer from a more academic setting to drug development practice, we also further develop the methodology to increase computational efficiency in practical settings. As such, our work enables quantitative scientist to transform (clinical) data into knowledge quicker and, crucially, in a replicable manner.

Wednesday, 06/Sept/2023 12:00pm - 12:20pm
ID: 199 / **S54: 5**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Software Engineering, Free Contributions
Keywords: Programming, reporting, multi-platform, collaboration, PDR technology

A multi-platform PDR technology to efficiently build complex tables & reports, flexibly iterate with stakeholders, and easily maintain across workflows

Ming Zou[1,2]
[1]RepTik Analytics Solution, Switzerland; [2]University Hospital of Basel, Switzerland; mzou@reptik.swiss

Background: Current programming techniques for customized data analysis and reporting, e.g., a clinical study report for regulatory stakeholders, are mostly via a "brute force" approach and have little breakthrough for the last 30 years. Programmers are wasting lots of recurrent efforts on supportive jobs like format/layout/shell programming and associated debug, QC, modify, manual edits, and trainings. Due to the limitation, programmers are reluctant to iterate with report stakeholders for changes. Further, it leads to highly fragmented and constrained practices across different teams/platforms/file types, which is very difficult to collaborate.

Method: We seek out to address the problems by first analyzing different teams' analysis and reporting workflows across multiple platforms (SAS, R, SQL…) and across multiple report file types (Word, Excel, PowerPoint…). Based on the analysis we designed and developed a multi-platform Presentation Data Referencing (PDR) technology, which enables an efficient, flexible, and inter-operable approach to perform customized analysis and reporting for clinical studies.

Result: From the workflow analysis, we discovered that due to limitations in current reporting techniques, data analysts typically tend to narrow themselves to one preferred report file type and one preferred analysis and scripting software, and then to maximize the efficiency by simply wrap up complex and constrained codes into big functions. For new situations, the recurrent iterative cycles of scripting, debugging, modifying, and quality checking activities cost a lot of efforts. Meanwhile, the value adding jobs are mainly the data analysis and result generating parts, not the format and layout generating parts. Inspired by using an index structure to operate numeric arrays, we designed and developed a multi-platform Presentation Data Referencing (PDR) technology by referencing the result placeholders in a report template (e.g. a Word shell report) and referencing the specific results during the analysis, so that to inject the values directly into the report template without programming it. In this way, users can create, format, and modify the tables & reports with the ease of office software and fill in with data from different platforms automatically through the PDR technology.

Conclusion: Our multi-platform PDR technology is applicable across different programming platforms, different analysis teams, and different report file types. Our technology can boost productivity, collaboration, and inter-operability of clinical studies for different organizations and stakeholders. The PDR technology is now patent-pending.

Thursday, 07/Sept/2023 11:00am - 11:20am
ID: 396 / **S64: 2**
Presentation Submissions - Regular Session (Default)
Topic of Submission: Estimands and causal inference
Keywords: Standard Endpoints, Clinical Interpretation, Lymphoma, Event Free Survival

Estimands in practice: revisiting a standard endpoint definition in the light of the lymphoma patient journey.

Emmanuel Zuber
Novartis, Switzerland; emmanuel.zuber@novartis.com

Event Free Survival (EFS) is considered a standard endpoint in Diffuse Large B-Cell Lymphoma (DLBCL). It measures the time to a composite event including tumor relapse, death or treatment failure, whichever occurs first. The first two types of events are traditionally well defined and not controversial in their ability to capture a detrimental outcome for the patient. On the other hand, treatment failure appears defined in very diverse ways across published trials, usually based on either the failure to reach complete or partial tumor response (CR/PR) by a given milestone, and/or on the administration of a further anticancer therapy. The contribution of this type of event to an objective assessment of treatment effect, comparatively to the two other types, appears harder to understand.

The review of a typical patient journey as referred to in the treatment guidelines helps to clarify the original clinical rationale for the definition of treatment failure in EFS. Achieving CR or PR has historically been a critical pre-requisite on the path to possible cure, in particular to enable the administration of high dose chemotherapy and stem cell transplantation (HDC/SCT).

However, the high diversity of implementation of the EFS endpoint across clinical studies, particularly in the definition of treatment failure, often blurs the relation to this original clinical rationale, and often doesn't come with an explicit clinical explanation. This is even more challenging when the EFS endpoint is used in new drug development settings, e.g., with novel treatment approaches not relying on HDC/HCT. When different treatment modalities need to be compared, the definition of treatment failure may raise further difficulties in the statistical analysis and interpretation of study results.

In this presentation, we review possible patient journeys and trial situations to highlight how various definitions of EFS may address different questions of interest. This highlights how the definition of this endpoint would benefit from a structured and transparent estimand discussion according to ICH E9(R1), centered around explicit clinical assumptions and intercurrent events grounded into patient journeys. This is essential to identify the treatment effect(s) of interest and guide the trial design and statistical analysis, to ensure interpretability of the trial results. This presentation also highlights the need for statisticians to engage early on with clinical partners, emphasizing the importance of understanding the clinical rationale for endpoint conventions through the estimand framework.

Monday, 04/Sept/2023 4:30pm - 4:50pm
ID: 434 / **S17: 2**
Presentation Submissions - Topic Contributed Session
Topic Contributed Session: Prognostic and predictive biomarkers in personalized medicine
Keywords: in-vitro drug screen, drug synergy, molecular biomarker, variable selection

Bayesian hierarchical models for biomarker discovery in drug combination screens

Manuela Zucknick
University of Oslo, Norway; manuela.zucknick@medisin.uio.no

With high-throughput drug sensitivity screens we can quickly test compounds on cancer cell lines to determine treatment efficacy. Since molecular characterisation of the cell lines by various omics data sets is frequently available, we can link molecular features to treatment efficacy. The estimation of drug synergy is important when testing multiple compounds, but in vitro cell viability measurements can be imprecise due to measurement errors, especially for drug combination experiments. To address this, we propose a Bayesian hierarchical model that uses our recently developed Bayesian model for synergy estimation with uncertainty quantification. The model accounts for synergy estimation uncertainty and selects promising biomarkers for future analysis using a horseshoe prior. Because of the typically low sample size, there is often not enough signal to decisively escape the global shrinkage in the horseshoe prior. To address this issue, we use a variable selection approach called Signal Adaptive Variable Selector to separate the posterior selection probability of a molecular feature from its conditional effect size, i.e. from the posterior mean of its coefficient conditional on it being selected. We demonstrate the model in an application on a large high-throughput dataset of melanoma cell lines. Joint work with Leiv Rønneberg, Pilar Ayuda-Durán, Sigve Nakken, Eivind Hovig, Robert Hanes, Aram Andersen, Tine Norman Alver, Jorrit Enserink and Paul Kirk.