21-25 August 2022

43rd Annual Conference of the International Society of Biostatistics

www.iscb2022.info

Programme
& Book of Abstracts

ISCB43
NEWCASTLE, UK | 2022

The Local Organising Committee of the 43rd Conference of the International Society for Clinical Biostatistics (ISCB) would like to thank its sponsors for their support.

Host Partner:

Newcastle University

Gold sponsor:

Roche

Silver sponsors:

STATA

timberlake
data science | statistics | econometrics

janssen | PHARMACEUTICAL COMPANIES OF Johnson & Johnson

Bronze sponsor:

eLife

Welcome
FROM SPC CHAIR

On behalf of the Scientific Programme Committee (SPC), it is my great pleasure to welcome you to the 43rd Annual Conference of the International Society for Clinical Biostatistics (ISCB). The SPC members have been working hard to put together an exciting scientific program that covers the extraordinary breath of clinical biostatistics.

On Sunday, four exciting pre-conference short-courses start the conference off. Two full-day courses focusing on statistical methods: *An introduction to machine learning for health research* by Paul Kirk and Leiv Rønneberg (both University of Cambridge) and *Design and Analysis of Precision Medicine trials* by James Wason (University of Newcastle) and Haiyan Zheng (University of Cambridge) are complemented by two half-day courses on computational aspects: *Statistical integration of multiple omics datasets using OmicsPLS* by Said el Bouhaddani, (UMC Utrecht) and Jeanine Houwing-Duistermaat (University of Bologna) and an *Introduction to the tidyverse* by John Thompson (University of Leicester).

Two highly distinguished speakers will present during the plenary sessions. The President's Invited Speaker John Carlin (University of Melbourne), and Keynote Speaker Bhramar Mukherjee (University Michigan) will surely provoke interesting discussions by contrasting the good and the bad of methods old (regression) and data new (Electronic Health Records).

I am particularly excited that one of our invited sessions will honour this year's student award recipients giving the leaders of tomorrow a chance to shine today. The other six invited sessions will focus on various hot topics in biostatistics namely: *Machine learning with small data, Master protocols, statistical simulation, ageing, Bayesian nonparametrics and ML for causal inference, Semi-competing risks and causal inference*.

On the last day, the *Early Career Biostatisticians' (ECB) Day* will give young researcher the opportunity to share experiences while the two mini-symposia *STRATOS* (STRengthening Analytical Thinking for Observational Studies) and *Modern software tools for modern statistics* will focus on statistical and computational aspects, respectively.

Complementing the arranged part of the scientific programme is a fantastic range of contributed talks and posters that show a fabulous mix of areas and ideas. Your contributions made the choice of what to include into the programme a difficult, but enjoyable, task and I am looking forward to learn about all the new insights and findings and will make sure to dip into areas not so familiar to me.

Thomas Jaki
Chair of ISCB43 Scientific Programme Committee

Welcome
FROM LOC CHAIR

On behalf of the Local Organizing Committee (LOC), I am delighted to welcome you to Newcastle for the 43rd Annual Conference of the International Society for Clinical Biostatistics (ISCB). This is the first hybrid conference for ISCB and we have worked hard to make the conference accessible, interactive and enjoyable for both online and in-person attendance. The conference takes place in the Frederick Douglass Centre, which opened in 2020, and is named in honour of social reformer, abolitionist and activist-author Frederick Douglass.

In-person attendees will discover that Newcastle has a compact city centre, with shops, restaurants, hotels and the beautiful quayside all within a short walk of the conference centre. Within the programme we have kept the Tuesday afternoon and evening free for you to explore the area and socialise with friends and colleagues old and new.

From the Angel of the North to the Tyne Bridge, Newcastle has much to offer. Explore the quayside on foot and enjoy the dramatic views of the seven bridges and iconic buildings. Experience the Roman history by visiting the UNESCO World Heritage site Hadrian's Wall. Contemporary art lovers should check out the Baltic, and bookworms will love the Lit & Phil, Newcastle's compact and exquisite independent library. If football is your thing book a tour of St James Stadium, walking distance from the Conference centre. Those young, or young at heart will enjoy a night out in Newcastle, which has a large selection of pubs, clubs and other entertainment.

A short journey on public transport will yield even more delights, Durham Castle and Cathedral (UNESCO World Heritage site), a traditional British seaside at Whitley Bay and castles and beautiful countryside throughout the region.

For those attending online we are streaming the full scientific programme live through our conference platform. While the in-person delegates are enjoying the reception on Monday we will be running a speed networking social via the online conference platform. You will also be able to contact and message other delegates though the app. On the Wednesday evening we have an online local history talk which will be open to all delegates.

The membership of the LOC includes biostatisticians from UCB, Johnson and Johnson, the Newcastle University Biostatistics Research Group and the School of Mathematics, Statistics and Physics. We look forward to hearing your presentations, viewing your posters and engaging in discussions on the hot topics for clinical biostatistics in 2022.

Dawn Teare
Chair of ISCB43 Local Organising Committee

Table of contents

Table of contents

Organisation – Committees

Scientific programme committee

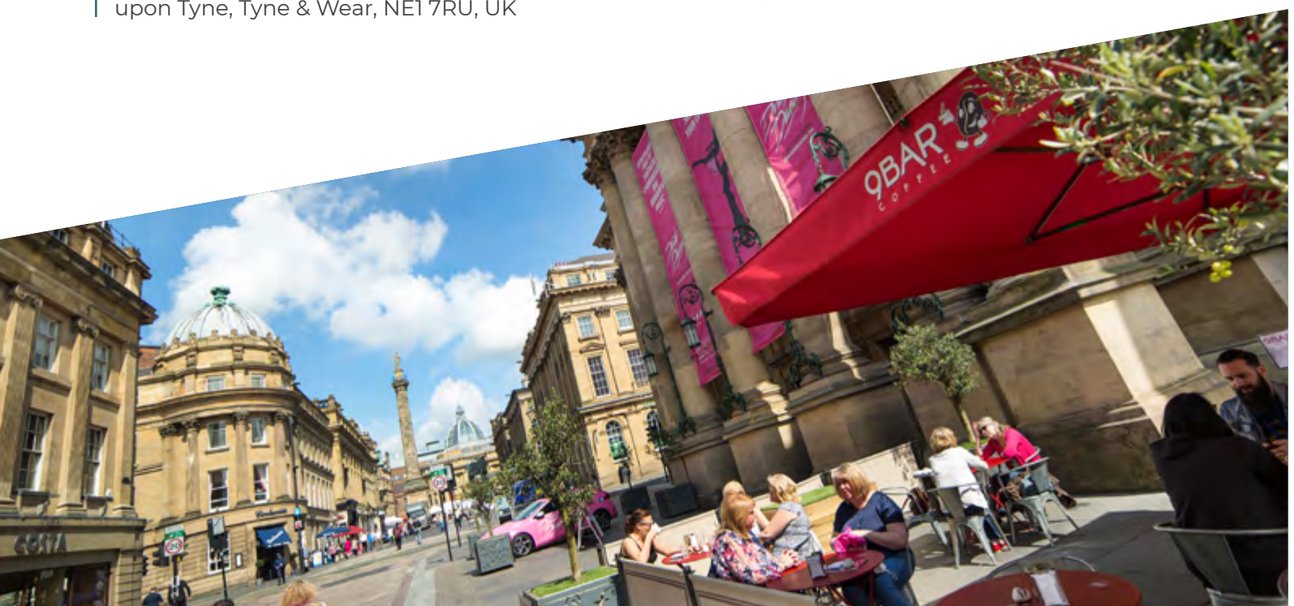
Thomas Jaki (UK, chair)
Gianluca Baio (UK)
Harald Binder (GE)
Svetlana Cherlin (UK)
Richard J. Cook (CA)
Michael J. Crowther (SE)
Michael J. Daniels (US)
Karla Diaz-Ordaz (UK)
Nan van Geloven (NL)
Satoshi Hattori (JP)
Leonhard Held (CH)
Jeanine Houwing-Duistermaat (UK)
Kelly Kidwell (US)
Oleguer Plana-Ripoll (DK)
Nathalie Støer (NO)
Ioanna Tzoulaki (GR)
An Vandebosch (BE)

Local Organising Committee

Dawn Teare | Chair
John Matthews | Treasurer
Svetlana Cherlin
Mike Cole
Michael Grayling
Adetayo S. Kasim
Jérémie Nsengimana
Luke Ouma
Vicky Ryan
James Wason
Faye Williamson
Nina Wilson

Organisation

Newcastle University, Newcastle upon Tyne, Tyne & Wear, NE1 7RU, UK



About ISCB

The International Society for Clinical Biostatistics (ISCB) was founded in 1978 to stimulate research into the principles and methodology used in the design and analysis of clinical research and to increase the relevance of statistical theory to the real world of clinical medicine.

Membership is open to all interested individuals who share the Aims of the Society. ISCB's membership includes clinicians, statisticians and members of other disciplines, such as epidemiologists, clinical chemists and clinical pharmacologists, working or interested in the field of clinical biostatistics.

ISCB has an Executive Committee and 6 Subcommittees: Conference Organising, Early Career Biostatisticians, Education, National Groups, Statistics in Regulatory Affairs (SiRA), Student Conference Awards.

Executive Committee

President | Zdenek Valenta (CZ)
Vice-President | Tomasz Burzykowski (BE)
Secretary | Elaine Pascoe (AU)
Treasurer | Chris Metcalfe (UK)

Members

Dimitris Rizopoulos (NL)
Gerta Ruecker (DE)
Jonathan Bartlett (UK)
Krystyna Szafraniec (PL)
Laure Wynants (NL)
Lehana Thabane (CA)
Nan van Geloven (NL)
Nicole Close (US)
David W. Warne (CH)

Honorary Members

Michal Abrahamowicz
Harbajan Chadha-Boreham
Lutz Edler
Emmanuel Lesaffre
Michael Schemper
Martin Schumacher
Jørgen Seldrup
Stephen J. Senn
Hans C. van Houwelingen
David W. Warne
John R. Whitehead

ISCB Permanent Office

convin

29 K. Varnali Street, 152 33
Chalandri,
Athens, Greece
Call center | +30 210 6833600
Mobile contact | +30 6956 665669
Website | www.iscb.info
Email | secretariat@iscb.info

2022 Award Recipients

Student Award Winners	Title of paper	Session code	University/Organisation	Country
Fangya Mao	**IN5.1** Spatial Dependence Modeling of Latent Susceptibility and Time to Joint Damage in Psoriatic Arthritis	**IN5.1**	University of Waterloo	Canada
Anthony Devaux	**IN5.4** Random survival forests for competing causes with multivariate longitudinal endogenous covariates	**IN5.4**	Université de Bordeaux	France
Theodoros Evrenoglou	**IN5.2** Sharing information across patient subgroups to draw conclusions from sparse treatment networks	**IN5.2**	Université Paris Cité	France
Laura Guedemann	**IN5.3** Triangulating Instrumental Variable, confounder adjustment and Difference-in-Difference methods for comparative effectiveness research in observational data	**IN5.3**	University of Exeter	United Kingdom
Science Award Winners	**Title of paper**	**Session code**	**University/Organisation**	**Country**
Suborna Sultana	**P71** Disagreement Based Variable Selection Method for High-dimensional Censored Data	**P71**	Dhaka University	Bangladesh
Saifur Rahman Mazumder	**P100** Implementation of Lightgbm and XGBoost to Survival Data under AFT Model	**P100**	Institute of Statistical Research and Training, University of Dhaka	Bangladesh
CFDC	**Title of paper**	**Session code**	**University/Organisation**	**Country**
Abdul Basit	**S2.2** Sensitivity Analysis of Causal Effects in Observational Studies	**S2.2**	Institute of Statistical Research & Training, University of Dhaka	Bangladesh
Baoyi Haung	Machine learning performs better in individual dynamic prediction with non-proportional hazards	**P112**	Southern Medical University	China
Chengfeng Zhang	**S4.4** Restricted mean survival time regression model with time-dependent covariates	**S4.4**	Southern Medical University	China
Afsana Dil Afroze	**P24** Analysis of Clustered Survival Data with Dependent Censoring	**P24**	Institute of Statistical Research and Training, University of Dhaka	Bangladesh
Oludare Ariyo	**S7.3** Prior distributions for structured covariance matrices	**S7.3**	Federal University of Agriculture, Abeokuta	Nigeria
Naser Kamyari	**P34** Bayesian Inference on Multilevel Semi-continuous Pharmaceutical expenditure Data; **P70** Diet, Nutrition, Obesity, and Their Implications for COVID-19 Mortality: A Marginalized 2-Part Model	**P34; P70**	Abadan University of Medical Sciences	Iran
Trinh Dong	**P13** A novel scoring system for diagnosis of Tuberculous Meningitis: a Bayesian Latent Class Analysis	**P13**	Oxford University Clinical Research Unit	Vietnam
Andrine Nyesiga	**P8** A Curve of healthy ageing found in older adults, A cross-sectional study on the oddity of aging.	**P8**	Uganda Development and Health Associates	Uganda
Emama Amin	**P4** A comparison of different estimands of causal effects in competing risks setting using a simulation	**P4**	Institute of Statistical Research and Training, University of Dhaka	Bangladesh
Robabeh Ghodssi-Ghassemabadi	**P179** Time-varying frailty multistate model in the presence of time-dependent covariates	**P179**	Tarbiat Modares University	Iran

Keynote Speakers

President's Invited Speaker

Prof John B. Carlin
Murdoch Children's Research Institute (AU)

Monday 22 August 2022 | 16.00-17.00
On the many uses and abuses of regression models

Keynote Invited Speaker

Prof Bhramar Mukherjee
University of Michigan School of Public Health (US)

Wednesday 24 August 2022 | 11.00-12.00
Using Electronic Health Records for Scientific Research: Promises and Perils

SUNDAY | 21 August 2022

PRE-CONFERENCE COURSES

	Room 2.15	Room 2.16	Room 2.14	Room 2.14
09.00-10.30	**Full day Shortcourse 1:** Introduction to Machine learning for health questions *Dr Paul Kirk*, *University of Cambridge*	**Full day Shortcourse 2:** Design and Analysis of Precision Medicine trials *Prof James Wason*, *Newcastle University, and Haiyan Zheng, University of Cambridge*	**Half day Shortcourse 3:** Statistical integration of multiple omics datasets using OmicsPLS *Said el Bouhaddani*, *UMC Utrecht and Jeanine Houwing-Duistermaat, University of Bologna*	
10.30-11.00	BREAK			
11.00-12.30	**Full day Shortcourse 1:** Introduction to Machine learning for health questions	**Full day Shortcourse 2:** Design and Analysis of Precision Medicine trials	**Half day Shortcourse 3:** Statistical integration of multiple omics datasets using OmicsPLS	
12.30-13.30	LUNCH			
13.30-15.00	**Full day Shortcourse 1:** Introduction to Machine learning for health questions	**Full day Shortcourse 2:** Design and Analysis of Precision Medicine trials		**Half day Shortcourse 4:** Introduction to the tidyverse *Prof John Thompson*, *University of Leicester*
15.00-15.30	BREAK			
15.30-17.00	**Full day Shortcourse 1:** Introduction to Machine learning for health questions	**Full day Shortcourse 2:** Design and Analysis of Precision Medicine trials		**Half day Shortcourse 4:** Introduction to the tidyverse

MONDAY | 22 August 2022

	Room G.41	Room G.56	Room G.06	Room 1.17	Room 2.16
09.00-10.30	**Invited 1:** Machine learning with small data	**Parallel Session 1:** Adaptive Designs	**Parallel Session 2:** Causal Inference	**Parallel Session 3:** Missing data	**Parallel Session 4:** Survival data
10.30-11.00	BREAK				
11.00-12.30	**Invited 2:** Master protocols	**Parallel Session 5:** Software Engineering	**Parallel Session 6:** Cluster Trials	**Parallel Session 7:** High dimensional data	**Parallel Session 8:** Lightning poster talks 1
12.30-13.30	LUNCH				
13.30-15.00	**Invited 3:** Best practices in statistical simulation and computing	**Parallel Session 9:** Meta-analysis	**Parallel Session 10:** Ageing	**Parallel Session 11:** Communication of survival methods	**Parallel Session 12:** Lightning poster talks 2
15.00-15.30	BREAK				
15.30-16.00	**Conference opening**				
16.00-17.00	**President's invited speaker:** Prof John Carlin				
17.00-19.00	17:00-19.00: **Welcome reception** (At conference venue) 17.15-18.00: **Networking social for virtual attendees (online)**				

TUESDAY | 23 August 2022

	Room G.41	Room G.56	Room G.06	Room 1.17	Room 2.16
09.00-10.30	**Invited 4:** Ageing	**Parallel Session 13:** COVID-19	**Parallel Session 14:** Basket trials	**Parallel Session 15:** Machine learning for health applications	**Parallel Session 16:** Lightning poster talks 3
10.30-11.00	BREAK				
11.00-12.30	**Invited 5:** Student Awardees	**Parallel Session 17:** Competing risks	**Parallel Session 18:** Personalized Medicine	**Parallel Session 19:** Clinical Trials	**Parallel Session 20:** Heterogeneity in effects
12.30-13.30	LUNCH				
13.30-17.00	**Social Excursions:** City walk around Newcastle Gateshead Meet at Grey's Monument at 14.00 Victoria tunnels tour Meet at Ouseburn for the tour Hadrian's wall leaving from the Frederick Douglass Centre at 13.30 These are all pre-bookable online (subject to availability): **https://www.iscb2022.info/programme/ Social Excursions page.** Ask for details about any places available at the registration desk.				

WEDNESDAY | 24 August 2022

	Room G.41	Room G.56	Room G.06	Room 1.17	Room 2.16
09.00-10.30	**Invited 6:** Bayesian nonparametrics and ML for causal inference	**Parallel Session 21:** Communicating statistical concepts	**Parallel Session 22:** Counterfactuals	**Parallel Session 23:** Bias and Estimation	**Parallel Session 24:** Lightning never strikes twice
10.30-11.00	BREAK				
11.00-12.00	**Keynote Speaker:** Prof Bhramar Mukherjee				
12.00-13.30	LUNCH & 12.15-13.15 AGM (room G.06)				
13.30-15.00	**Invited 7:** Semi-competing risks and causal inference	**Parallel Session 25:** Simulation and software	**Parallel Session 26:** Missing data in Studies	**Parallel Session 27:** Prediction modelling	**Parallel Session 28:** Efficient trial designs
15.00-15.30	BREAK				
15.30-17.00	**Parallel Session 29:** Joint inference with high dimensional data	**Parallel Session 30:** Machine learning and prediction	**Parallel Session 31:** Early Phase Trials	**Parallel Session 32:** Complex Modelling	
19.00-23.00	**Conference Dinner** Civic Centre Newcastle 19.00-19.45: **History talk: Meet the story of Coal to Culture for virtual attendees (online)**				

Programme overview

THURSDAY | 25 August 2022

	Room G.06	Room G.41	Room G.56
09.00-10.30	**Mini symposium 1:** STRATOS	**Mini symposium 2:** Modern software tools for modern statistics	**ECB Day**
10.30-11.00	BREAK		
11.00-12.30	**Mini symposium 1:** STRATOS	**Mini symposium 2:** Modern software tools for modern statistics	**ECB Day**

Detailed Programme

ISCB43
NEWCASTLE,UK | 2022

ROOM 2.15

09:00-17:00 PRE-CONFERENCE COURSE 1: Introduction to Machine Learning for health questions

FULL DAY COURSE

Presenter: **Paul Kirk**, *University of Cambridge, UK*

TIMELINE:

09:00-10:30	▸	Part A
10:30-11:00	▸	BREAK
11:00-12:30	▸	Part B
12:30-13:30	▸	LUNCH
13:30-15:00	▸	Part C
15:00-15:30	▸	BREAK
15:30-17:00	▸	Part D

ABSTRACT:
In this one-day short course, we will provide a broad grounding in machine learning basics, illustrated throughout by applications in biohealth. Methodological topics will include: (i) Modelling, generalisation, and overfitting; (ii) An introduction to regression and classification; (iii) Unsupervised learning: clustering, density estimation, and dimension reduction; and (iv) Advanced predictive modelling: artificial neural networks and Gaussian processes. Topics will be accompanied throughout by practical illustrations and applications in R.
Presenters: Paul Kirk is an MRC Investigator at the MRC Biostatistics Unit, University of Cambridge, whose research group uses Bayesian statistics and machine learning to address questions in molecular precision medicine and to identify patterns in electronic health record databases.

ROOM 2.16

09:00-17:00 PRE-CONFERENCE COURSE 2: Design and Analysis of Precision Medicine trials

FULL DAY COURSE

Presenters: **James Wason**, *Newcastle University, UK*
Haiyan Zheng, *University of Cambridge, UK*

TIMELINE:

09:00-10:30	▸	Part A
10:30-11:00	▸	BREAK
11:00-12:30	▸	Part B
12:30-13:30	▸	LUNCH
13:30-15:00	▸	Part C
15:00-15:30	▸	BREAK
15:30-17:00	▸	Part D

ABSTRACT:
Precision medicine is about going beyond assessing whether a new treatment works on average to predicting which subgroups of patients receive benefit and to what extent. When the subgroups, often defined by biomarkers, genetic, phenotypic or psychosocial characteristics, are associated with a treatment's efficacy or toxicity, precision medicine offers substantial advantages to patients, trial sponsors, and the wider healthcare system.
In this course we introduce the concept of precision medicine and cover some innovative trial design and analysis approaches, including basket, umbrella, Bayesian hierarchical modelling, adaptive signature and adaptive enrichment designs. These approaches have all been developed to improve power and patient benefit provided by clinical trials. Examples from a wide variety of therapeutic areas will be discussed, with implementation in OpenBUGS and R software. Perspectives will be given on the future development of design, conduct and analysis of clinical trials in the field.

SUNDAY 21 AUGUST 2022

ROOM 2.14

09:00-12:30 PRE-CONFERENCE COURSE 3: Statistical integration of multiple omics datasets using OmicsPLS

HALF DAY COURSE

Presenters: **Said el Bouhaddani**, *UMC Utrecht, NL*
Jeanine Houwing-Duistermaat, *University of Bologna, IT*

TIMELINE:

09:00-10:30	▸	Part A
10:30-11:00	▸	BREAK
11:00-12:30	▸	Part B

ABSTRACT:
In many epidemiological and clinical studies, multiple omics datasets are available, e.g. transcriptomics, glycomics, methylation. Since these datasets are supposed to represent the same complex biological mechanisms, data integration methods are typically applied. Two main challenges arise when dealing with omics data integration: (i) high dimensional and highly correlated features, (ii) heterogeneity among omics data. Partial least squares (PLS) and its extension two-way orthogonal PLS (O2PLS) address these challenges. These methods extract linear components which represent the relationship between the datasets (dimension reduction) and identify the most relevant features explaining this relationship. To facilitate interpretation of relevant features, a sparse group O2PLS approach can be used (GO2PLS). These methods are implemented in the open-source 'OmicsPLS' R package.
This half-day course introduces the fundamental concepts behind omics data integration based on latent variable models and joint principal components. We will explain how to deal with heterogeneity between omics data and how to perform feature selection. Emphasis is given on applications, and the OmicsPLS package will be introduced and used during a practical R session. At the end of the course, participants should be able to perform and interpret omics data integration with OmicsPLS.

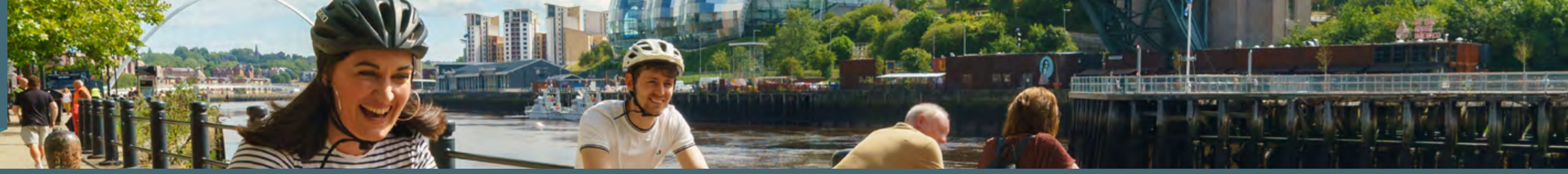
OUTLINE:
The course consists of two parts. The first session is a mix of theory and practice. You will be introduced to the basics of data integration approaches and their sparse variants. We will also discuss how to incorporate group structures such as CpG sites of a gene and genes of a pathway. You will practice with these approaches using OmicsPLS and try out different visualisation tools to interpret the results. The second breakout session will be hands-on, where you apply what you've learned to two omics datasets. There will be a brief rejoinder at the end of the session to discuss the output of the exercises.

1) Dimension reduction with PCA
2) Essentials of data integration: Partial least squares (PLS) approach
3) Multi-omics data integration with O2PLS
4) Methods to incorporate external biological information: Sparse group O2PLS (GO2PLS)
5) Post-hoc analyses of the results using external bioinformatics databases

By the end of this course, you should be able to:

- Have a deeper understanding of data integration methods (PLS, O2PLS, GO2PLS)
- Understand how feature selection works
- Implement O2PLS/GO2PLS on your own data with "OmicsPLS" R package
- Interpret and visualise O2PLS/GO2PLS results

TARGET AUDIENCE:
The course is aimed at biostatisticians or medical researchers working with (high-dimensional) biological data, including multi-omics data, and want to learn how to combine these data. Participants are expected to be familiar with linear regression modelling, principal component analysis and have basic knowledge in R.



SUNDAY 21 AUGUST 2022

ROOM 2.14

13:30-17:00 PRE-CONFERENCE COURSE 4: Introduction to the tidyverse

HALF DAY COURSE

Presenter: **John Thompson**, *University of Leicester, UK*

TIMELINE:

13:30-15:00	▸	Part A
15:00-15:30	▸	BREAK
15:30-17:00	▸	Part B

OUTLINE:
The pipe operator, %>%, greatly simplifies statistical programming in R. The tidyverse is a collection of R packages that are designed to work with the pipe to produce concise, readable and very efficient R code. The tidyverse is particularly suited to data analysis and modelling. This course will introduce the tidyverse by taking a dataset and processing it, from reading the data through to the final report. The course will emphasise the three most important tidyverse packages; dplyr, a package for data manipulation, ggplot2, a package for graphics and rmarkdown, a package for report production. Time will be set aside for the course participants to run their own analysis of a similar dataset. The course will end with a brief overview of the more advanced tidyverse packages and their associated programming styles, including the map functions from the purrr package. The course is suitable for someone who has a basic knowledge of R and RStudio and who wants to modernise and extend their skills. If possible, bring along your own laptop with R and RStudio installed.
John is a biostatistician with an interest in epidemiology and the analysis of ge-netic studies. He was Professor of Genetic Epidemiology at the University of Leicester until he retired just before COVID struck. He now has an Emeritus post at the University. Over the last decade, John experimented with a range of ways of teaching R before settling on the tidyverse and the problem-based ap-proach that will be used for this short course. If you want to know more about John's approach to teaching R, he has a series of blog posts on the subject that start here.

MONDAY 22 AUGUST 2022

ROOM G.41

09:00-10:30 INVITED SESSION 1: Machine learning with small data

Chair: **Svetlana Cherlin** (UK)

09:00-09:30	**IN1.1**	**Manuela Zucknick** (NO)	*Structured priors for improving treatment response prediction in cancer pharmacogenomic screens*
09:30-10:00	**IN1.2**	**Wessel van Wieringen** (NL)	*Sequential learning of regression models through penalized estimation*
10:00-10:30	**IN1.3**	**Moritz Hess** (DE)	*Interpretable deep generative approaches for small omics data-sets*

ROOM G.56

09:00-10:30 PARALLEL SESSION 1: Adaptive designs

Chair: **Franz Koenig** (AT)

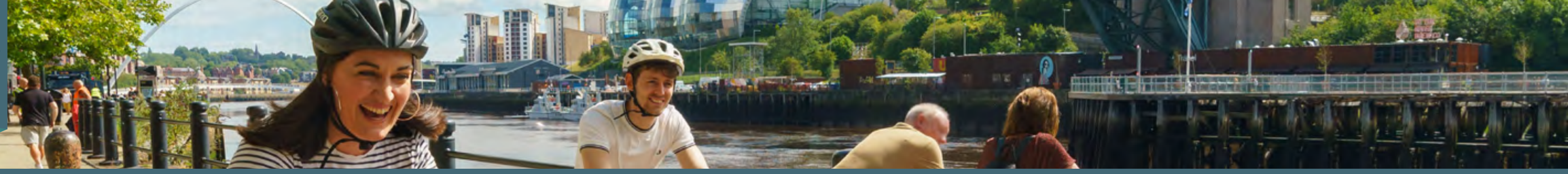
09:00-09:18	**S1.1**	**Nigel Stallard** (UK)	*Adaptive enrichment designs with a continuous biomarker*
09:19-09:36	**S1.2**	**Alessandra Serra** (UK)	*A Bayesian multi-arm multi-stage design incorporating information about treatment ordering*
09:37-09:54	**S1.3**	**Akane Yamakawa** (JP)	*Recommending a timing for a stop for efficacy in group sequential trials with a survival endpoint*
09:55-10:12	**S1.4**	**Nico Bruder** (DE)	*Two-stage designs for clinical trials with small sample sizes*
10:13-10:30	**S1.5**	**Svetlana Cherlin** (UK)	*Cross-validated risk scores adaptive enrichment design*

ROOM G.06

09:00-10:30 PARALLEL SESSION 2: Causal Inference

Chair: **Saskia Le Cessie** (NL)

09:00-09:18	**S2.1**	**Ashish Patel** (UK)	*Using many invalid instrumental variables to tighten inference on causal effects*
09:19-09:36	**S2.2**	**Md. Abdul Basit** (BD) ***CFDC recipient***	*Sensitivity Analysis of Causal Effects in Observational Studies*
09:37-09:54	**S2.3**	**Sarah Booth** (UK)	*Investigating inequalities in cancer survival through mediation analysis with limited interventions*
09:55-10:12	**S2.4**	**Sharon Lutz** (US)	*Inferring the Effect Direction in Genetic Association Studies*
10:13-10:30	**S2.5**	**Zhenwei Yang (NL)**	*Personalized Biopsy Schedules Using Cause-specific Interval-censored Joint Models*



ROOM 1.17

09:00-10:30 PARALLEL SESSION 3: Missing data

Chair: **Tim Morris** (UK)

09:00-09:18	**S3.1**	**Edouard F Bonneville** (NL)	*Multiple imputation for the Fine-Gray model: an approach based on the subdistribution process*
09:19-09:36	**S3.2**	**Michel Hof** (NL)	*Record linkage with complex correlation structures*
09:37-09:54	**S3.3**	**Emily Kawabata** (UK)	*Quantitative bias analysis for unmeasured confounding: Review of software for regression*
09:55-10:12	**S3.4**	**Li Su** (UK)	*Sensitivity analysis for calibrated weighted estimators under non-ignorable dropout*
10:13-10:30	**S3.5**	**Marcel Wolbers** (CH)	*Standard and reference-based imputation methods based on conditional mean imputation*

ROOM 2.16

09:00-10:30 PARALLEL SESSION 4: Survival data

Chair: **Geir Egil Eide** (NO)

09:00-09:18	**S4.1**	**Jasmin Rühl** (DE)	*General independent censoring in event-driven trials with staggered entry*
09:19-09:36	**S4.2**	**Kaspar Rufibach** (CH)	*Follow-up time in clinical trials with a time-to-event endpoint: Redefining the question(s)*
09:37-09:54	**S4.3**	**Dominic Magirr** (CH)	*Stratified modestly-weighted log-rank tests in settings with delayed separation of survival curves*
09:55-10:12	**S4.4**	**Chengfeng Zhang** (CN) **CFDC recipient**	*Restricted mean survival time regression model with time-dependent covariates*
10:13-10:30	**S4.5**	**Denis Rustand** (SA)	*Efficient estimation of joint models for multivariate longitudinal and survival data using INLA*

10:30-11:00 BREAK

ROOM G.41

11:00-12:30 INVITED SESSION 2: Master protocols

Chair: **Thomas Jaki** (UK)

11:00-11:22	**IN2.1**	**James Wason** (UK)	*Statistical issues in design and analysis of master protocols*
11:23-11:45	**IN2.2**	**Franz Koenig** (AT)	*Multiplicity Issues in Exploratory Platform Trials – Online Control of the False Discovery Rate*
11:46-12:07	**IN2.3**	**Emily Zabor** (US)	*Bayesian Basket Trial Design with False Discovery Rate Control*
12:08-12:30	**IN2.4**	**Mary Redman** (US)	*On the Lung Cancer Master Protocol (Lung-MAP)*

MONDAY 22 AUGUST 2022

ROOM G.56

11:00-12:30 PARALLEL SESSION 5: Software Engineering

Chair: **Oliver Boix** (DE)

11:00-11:18	**S5.1**	**Martin Shaw** (UK)	*Introduction from the Academic Perspective*
11:19-11:36	**S5.2**	**Daniel Sabanes Bove** (CH)	*Introduction from the Industry Perspective*
11:37-12:30	**S5.3**	**Armin Schueler** (DE) **Andy Nicholls** (UK) **Anne-Laure Boulesteix** (DE) **Alessandro Gasparini** (SE) **Martin Shaw** (UK) **Daniel Sabanes Bove** (CH)	**PANEL DISCUSSION**

ROOM G.06

11:00-12:30 PARALLEL SESSION 6: Cluster Trials

Chair: **Dawn Teare** (UK)

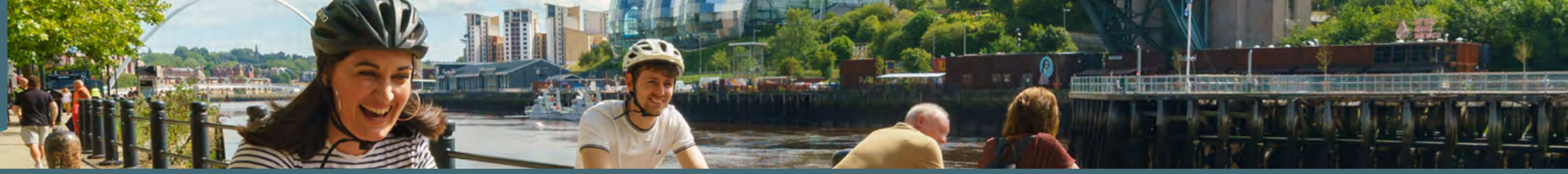
11:00-11:18	**S6.1**	**Jessica Kasza** (AU)	*Decaying correlation parameter values obtained from previously analysed cluster randomised trials*
11:19-11:36	**S6.2**	**Kelsey Grantham** (AU)	*The staircase cluster randomised trial design: a pragmatic alternative to the stepped wedge design*
11:37-11:54	**S6.3**	**Stephen Senn** (UK)	*What is the point of point estimates?*
11:55-12:12	**S6.4**	**Xiangmei Ma** (SG)	*Ratio-of-ratio estimator of direct intervention effect in cluster-randomized trials*
12:13-12:30	**S6.5**	**Floriane Le Vilain-Abraham** (FR)	*Estimating the intervention effect using restricted mean survival time in a cluster randomized trial*

ROOM 1.17

11:00-12:30 PARALLEL SESSION 7: High dimensional data

Chair: **Maren Hackenberg** (DE)

11:00-11:18	**S7.1**	**Emilie Eliseussen Ødegaard** (NO)	*Rank-based Bayesian variable selection for genome-wide transcriptomic analyses*
11:19-11:36	**S7.2**	**Laia Canal Guitart** (DE)	*More than meets the eye: Visualising temporal patterns in time-series single-cell RNA-seq data*
11:37-11:54	**S7.3**	**Oludare Samuel Ariyo** (NG) **CFDC recipient**	*Prior distributions for structured covariance matrices*
11:55-12:12	**S7.4**	**Wencan Zhu** (FR)	*Identification of prognostic and predictive biomarkers in high-dimensional data with PPLasso*
12:13-12:30	**S7.5**	**Mark van de Wiel** (NL)	*Fast marginal likelihood estimation of group-adaptive elastic net penalties*



ROOM 2.16

11:00-12:30 PARALLEL SESSION 8: Lightning poster talks 1

Chair: **Richard Emsley** (UK)

P6	**Jacqueline Thompson** (UK)	*A comparison of methods for estimating dichotomous treatment effects: a simulation study*
P7	**Rheanna Mainzer** (AU)	*A comparison of strategies for selecting auxiliary variables for multiple imputation*
P9	**Reuben Adatorwovor** (US)	*A Flexible Copula Model for Bivariate Survival Data with Dependent Censoring*
P10	**Mari Brathovde** (NO)	*A lean additive frailty model: with an application to clustering of melanoma in Norwegian families*
P11	**Peter Greenstreet** (UK)	*A multi-arm multi-stage platform design that allows pre-planned addition of arms while still control*
P13	**Trinh Dong** (VN)	*A novel scoring system for diagnosis of Tuberculous Meningitis: a Bayesian Latent Class Analysis*
P17	**Zhulin Yin** (UK)	*A review of early phase dose-finding clinical trials with incomplete follow-up for toxicity consider*
P25	**Michal Kahle** (CZ)	*Analysis of concentration-time-response data in cell-based compound profiling*
P34	**Naser Kamyari** (IR)	*Bayesian Inference on Multilevel Semi-continuous Pharmaceutical expenditure Data*
P37	**Pedro Cardoso** (UK)	*Bayesian nonparametric methods for prediction of missing data*
P39	**Rugare Maruzani** (UK)	*Benchmarking Six Variant Callers for Detecting Low Frequency Variants in Circulating Tumour DNA*
P42	**Marta Spreafico** (IT)	*Causal effects of chemotherapy dose intensity on survival outcome through Marginal Structural Models*
P43	**Eleanor Van Vogt** (UK)	*Causal forests for uncovering treatment effect heterogeneity and data driven subgroups in trials*
P49	**Anca Chisster** (UK)	*Combining non-adherence and mediation in a unified causal analysis*
P57	**Yirui Qian** (UK)	*Comparison of statistical methods for the analysis of SF-36 in RCTs: an empirical analysis*
P62	**Joris Menten** (BE)	*Correlates of Protection Analysis of Vaccines in a Pre-Exposed Population*
P1	**Lubomir Stepanek** (CZ)	*"Great in, great out": how to select a subpopulation from data with no response variable*
P31	**Sarwar Mozumder** (UK)	*Assessment of prognostic model performance: A competing risks cause-specific hazards approach*
P126	**Jan Terje Kvaløy** (NO)	*Monitoring time to event in medical registry data using CUSUMs based on excess hazard models*
P142	**David van Klaveren** (NL)	*Predictive modeling approaches to personalized medicine: a comparison of regression-based methods*

12:30-13:30 LUNCH

ROOM G.41

13:30-15:00 INVITED SESSION 3: Best practices in statistical simulation and computing

Chair: **Michael Crowther** (SE)

13:30-14:00	**IN3.1**	**Anne-Laure Boulesteix** (DE)	*On the researchers' degree of freedom in comparison studies*
14:00-14:30	**IN3.2**	**Tim P Morris** (UK)	*Neutral schmeutral: fair comparisons and simulation-study estimand*
14:30-15:00	**IN3.3**	**Ricardo Sanchez** (US)	*Step-by-step guidance on best practices in statistical computing*

ROOM G.56

13:30-15:00 PARALLEL SESSION 9: Meta-analysis

Chair: **Simon Day** (UK)

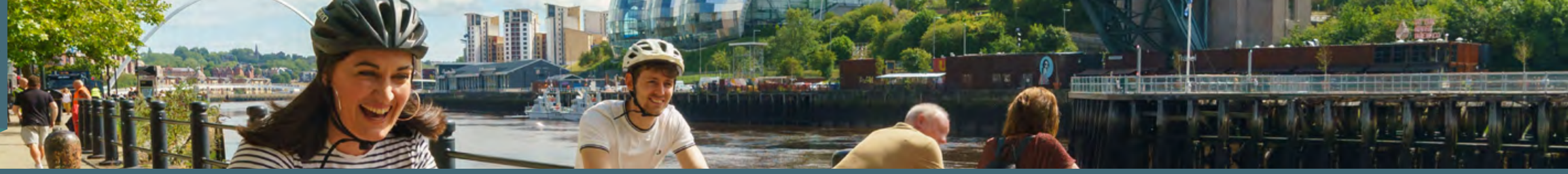
13:30-13:48	**S9.1**	**Elnaz Saeedi** (UK)	*A New Visualisation for Component Network Meta-Analysis: The Circle plot*
13:49-14:06	**S9.2**	**Stephen Walter** (CA)	*The usual method of estimating weights for studies in meta-analyses is biased*
14:07-14:24	**S9.3**	**Virginia Chiocchia** (CH)	*Ranking treatments on multiple outcomes and trade-off between benefit and harms*
14:25-14:42	**S9.4**	**David Fisher** (UK)	*Using cutting-edge methodology to maximise the value of Individual Participant Data meta-analysis*
14:42-15:00	**S9.5**	**Theodosia Salika** (UK)	*A Simulation Study Comparing Methods for Meta-Analysis of Time-to-Event Outcomes*

ROOM G.06

13:30-15:00 PARALLEL SESSION 10: Ageing

Chair: **Jeanne Houwing-Duistermaat** (UK)

13:30-13:48	**S10.1**	**Tiphaine Saulnier** (FR)	*4-step longitudinal analysis of latent traits derived from measurement scales in chronic diseases*
13:49-14:06	**S10.2**	**Marije Sluiskes** (NL)	*Biological age cannot be estimated with cross-sectional data*
14:07-14:24	**S10.3**	**Mizanur Khondoker** (UK)	*Multimorbidity pattern and risk of dementia: an 11-year follow-up study using the UK Biobank cohort*
14:25-14:42	**S10.4**	**Maja Pohar Perme** (SI)	*Expected life years compared to the general population*
14:42-15:00	**S10.5**	**Jiaxin Zhang** (AU)	*Recoverability and estimation of causal effects under typical multivariable missingness mechanisms*



ROOM 1.17

13:30-15:00 PARALLEL SESSION 11: Communication of survival methods

Chair: **Philip Hougaard** (DK)

13:30-13:48	**S11.1**	**Nikolaos Skourlis** (SE)	*Different multi-state structures when studying antidepressive medication in women with breast cancer*
13:49-14:06	**S11.2**	**Molly Wells** (UK)	*Fair comparison of cause-specific and relative survival by accounting for informative censoring*
14:07-14:24	**S11.3**	**Thanh Huan Vo** (FR)	*Cox regression with linked data*
14:25-14:42	**S11.4**	**Bor Vratanar** (SI)	*Evaluating cancer screening programmes using survival analysis*
14:42-15:00	**S11.5**	**Elisavet Syriopoulou** (SE)	*Assessing lead time bias due to mammography screening on estimates of loss in life expectancy*

ROOM 2.16

13:30-15:00 PARALLEL SESSION 12: Lightning poster talks 2

Chair: **Linda Sharples** (UK)

P65	**Roma Puronaite** (LT)	*Developing a Patient Engagement Framework for Quality Improvement in Healthcare Services*
P70	**Naser Kamyari** (IR)	*Diet, Nutrition, Obesity, and Their Implications for COVID-19 Mortality: A Marginalized 2-Part Model*
P72	**Giulia Barbati** (IT)	*Discrimination and calibration of atrial fibrillation predictions obtained by CNN and XGB from ECG*
P84	**Kim Luijken** (NL)	*Exploratory analyses in etiologic research and considerations for assessment of credibility*
P90	**Jing Haan-Du** (NL)	*Glycemic control prior to cancer incidence and mortality among patients with type 2 diabetes*
P91	**Ghazaleh Dashti** (AU)	*Handling multivariable missing data in causal mediation analysis with a single mediator*
P93	**Nuria Senar Villadeamigo** (NL)	*Hypothesis testing for CCA given prespecified sparsity levels*
P97	**Juliette Ortholand** (FR)	*Impact of ALS subtypes on disease progression: A continuous temporal multivariate approach*
P101	**Hannah Johnson** (UK)	*Improving Signal Detection in Small Pharmacovigilance Datasets – A New Pipeline Approach*
P102	**Johanna Munoz** (NL)	*Imputation of MNAR variables in an individual participant data meta-analysis*
P107	**Pierre-Emmanuel Poulet** (FR)	*Learning treatment effect in neurodegenerative diseases with a Bayesian mixed-effect model*
P108	**Phuc Tran** (IT)	*Likelihood-based inference in control risk regression with study-specific covariates*
P111	**Autumn O'Donnell** (IE)	*Machine learning methodologies for modelling of time-to-event endpoints in Prostate Cancer*
P115	**Jonathan Broomfield** (UK)	*Methods for modelling the multi-state natural history of rare diseases using disparate IPD sources.*
P119	**Marco Palma** (UK)	*Mixed-effects location-scale models for within-individual variability in cystic fibrosis*

P122	**Sida Chen** (UK)	*Modelling disease transitions in multimorbidity via multistate models*
P123	**Xiangmei Ma** (SG)	*Modelling non-linear time-varying intervention effects on recurrent events*
P166	**Gregor Buch** (DE)	*Sparse group penalties for bi-level variable selection*
P92	**Vincent Jeanselme** (UK)	*How do standard imputation strategies impact fairness under clinical presence?*
P139	**Even Moa Myklebust** (NO)	*Phenotypic deconvolution in heterogeneous cancer cell populations using drug screening data*
P48	**Maren Hackenberg** (DE)	*Combining deep learning and dynamic modelling to infer disease trajectories of patients with SMA*

15:00-15:30 BREAK

ROOM G.41

15:30-16:00 CONFERENCE OPENING

Welcome from Newcastle University | **Julie Sanders** (UK), *Deputy Vice-Chancellor*

Welcome from Conference Local Organising Committee Chair | **Dawn Teare** (UK)

Welcome from Conference Scientific Committee Chair | **Thomas Jaki** (UK)

Presentation of ISCB Conference Awards

ROOM G.41

16:00-17:00 PRESIDENT'S INVITED SPEAKER LECTURE: On the many uses and abuses of regression models

Introduction by: **Zdenek Valenta** (CZ), *ISCB President*

President's Invited Speaker: **Prof. John B. Carlin** (AU), *Murdoch Children's Research Institute*

CONFERENCE VENUE

17:00-19:00 WELCOME RECEPTION

ONLINE

17:15-18:00 NETWORKING SOCIAL PROGRAMME FOR ONLINE ATTENDEES

ROOM G.41

09:00-10:30 INVITED SESSION 4: Ageing

Chair: **Maja Pohar Perme** (SI)

09:00-09:25	**IN4.1**	**Andrew Clegg** (UK)	*Translating models of ageing into routine clinical practice*
09:25-09:50	**IN4.2**	**Mar Rodriguez-Girondo** (NL)	*A new statistical framework for biological age estimation*
09:50-10:15	**IN4.3**	**Joris Deelen** (DE)	*From variant to function: Identification and characterisation of genetic variants linked to human longevity*
10:15-10:30		**Jeanine Houwing-Duistermaat** (UK)	**DISCUSSION**

ROOM G.56

09:00-10:30 PARALLEL SESSION 13: COVID-19

Chair: **Steven Julious** (UK)

09:00-09:18	**S13.1**	**Hege Michiels** (BE)	*Estimation and interpretation of vaccine efficacy in COVID-19 randomized clinical trials*
09:19-09:36	**S13.2**	**Maiju Pesonen** (NO)	*The impact of nosocomial bacterial co-infections on the mortality of COVID-19 patients*
09:37-09:54	**S13.3**	**Micha Mandel** (IL)	*Covid-19: waning immunity and the booster dose effect in Israel*
09:55-10:12	**S13.4**	**Rachael Stannard** (UK)	*Assessing the impact of elevated population mortality rates due to Covid-19 on relative survival*
10:13-10:30	**S13.5**	**Vera Arntzen** (NL)	*Estimation of incubation time in relation to quarantine length: the impact of distributional assumptions*

ROOM G.06

09:00-10:30 PARALLEL SESSION 14: Basket trials

Chair: **Emily Zabor** (US)

09:00-09:18	**S14.1**	**Haiyan Zheng** (UK)	*Sample size determination in basket trials borrowing information across subsets*
09:19-09:36	**S14.2**	**Libby Daniells** (UK)	*Information Borrowing in Basket Trials: A Proposal and Evaluation*
09:37-09:54	**S14.3**	**Lou Whitehead** (UK)	*Bayesian information sharing methods for a longitudinal basket trial*
09:55-10:12	**S14.4**	**Luke Ouma** (UK)	*Bayesian modelling strategies for borrowing of information in randomised basket trials*
10:13-10:30	**S14.5**	**Lukas Baumann** (DE)	*Empirical Bayes Power Priors for Designing Basket Trials*

TUESDAY 23 AUGUST 2022

ROOM 1.17

09:00-10:30 PARALLEL SESSION 15: Machine learning for health applications

Chair: **Leiv Ronneberg** (UK)

09:00-09:18	**S15.1**	**Elvire Roblin** (FR)	*Uncertainty measures of survival predictions with neural networks applied to molecular data*
09:19-09:36	**S15.2**	**Martin Prodel** (FR)	*Subgroup discovery with survival decision trees: detection of early conversion in Alzheimer's stages*
09:37-09:54	**S15.3**	**Sonia Dembowska** (UK)	*Temporal Functional Principal Component Analysis*
09:55-10:12	**S15.4**	**Susan Ellul** (AU)	*Causal machine learning and use of sample splitting in settings with high-dimensional confounding*
10:13-10:30	**S15.5**	**Theophilus Quachie Asenso** (NO)	*Modeling high-dimensional interaction problems with the pliable lasso*

ROOM 2.16

09:00-10:30 PARALLEL SESSION 16: Lightning poster talks 3

Chair: **Charlotte Bolch** (US)

P124	**Natasa Kejzar** (SI)	*Modelling the impact of SARS-CoV-2 vaccination in a cohort of patients hospitalized for COVID-19*
P132	**Chikêola Ladekpo** (BE)	*Outlier study detection in a meta-analysis of clinical trials*
P134	**Jan Vávra** (CZ)	*Pattern Identification in Biomedical Markers of a Mixed Type*
P135	**Carolien Maas** (NL)	*Performance metrics for models predicting individualized treatment effect of patients*
P137	**Junfeng Wang** (NL)	*Performance of methods for meta-analysis of incremental predictive value: a comparison study*
P138	**Katie Scandrett** (UK)	*Performance of the P30 measure for assessing the accuracy of estimating glomerular filtration rate*
P140	**Natalia Pallares** (ES)	*Predicting multidrug resistance in neutropenic cancer patients with bloodstream infection*
P141	**Kleio Kipourou** (UK)	*Prediction of clinical trial cycle times with Bayesian Model Averaging*
P145	**Amelia Thompson** (UK)	*Quantitative decision-making in the context of early-phase biomarker-adaptive designs*
P151	**Nivetha Sridharan** (UK)	*Repeatability of Dynamic Contrast Enhanced Magnetic Resonance Imaging (DCE-MRI): A Systematic Review*
P152	**Theophile Bigirumurame** (UK)	*Reporting of sequential multiple assignment randomized trial design studies: a systematic review*
P157	**Bruno Pereira** (FR)	*Sample size estimation in RCT with annualized relapse rates can be improved: a review*
P162	**Sunita Rehal** (UK)	*Simulating the Impact of Intercurrent Events and Missing Data for Clinical Trials*
P163	**Simon Baldwin** (UK)	*Simulation study of within-subject variability estimation: linear mixed effects & variogram analyses*
P171	**Philip Hougaard** (DK)	*Survival of Danish twins born 1870-2000*
P175	**Kelly Grant** (UK)	*The Evaluation of Repeated Events in the PLEASANT dataset*
P176	**Pablo Emilio Verde** (DE)	*The Hierarchical Bias-Corrected Meta-Analysis Model*

P178	**Drifa Belhadi** (FR)	*The number of cases, mortality and treatments of viral hemorrhagic fevers: a systematic review*
P104	**Holly Tovey** (UK)	*Investigating use of Random Forest Clustering with multi-omics data to identify novel TNBC clusters*
P120	**Kaya Miah** (DE)	*Model selection for multi-state models in medical research*
P174	**Willi Sauerbrei** (DE)	*The effect of sample size on selecting a satisfactory descriptive model with the MFP approach*
P188	**Celina Gehringer** (UK)	*Multinomial prediction models for methotrexate outcomes in rheumatoid arthritis*
P164	**Miyui Hara** (JP)	*Simultaneous evaluation of superiority and non-inferiority*
P126	**Jan Terje Kvaløy** (NO)	*Monitoring time to event in medical registry data using CUSUMs based on excess hazard models*
P142	**David van Klaveren** (NL)	*Predictive modeling approaches to personalized medicine: a comparison of regression-based methods*

10:30-11:00 **BREAK**

ROOM G.41

11:00-12:30 **INVITED SESSION 5: Student Awardees**

Chair: **Jonathan Bartlett** (UK)

11:00-11:22	**IN5.1**	**Fangya Mao** (CA)	*Spatial Dependence Modeling of Latent Susceptibility and Time to Joint Damage in Psoriatic Arthritis*
11:23-11:45	**IN5.2**	**Theodoros Evrenoglou** (FR)	*Sharing information across patient subgroups to draw conclusions from sparse treatment networks*
11:46-12:07	**IN5.3**	**Laura Guedemann** (UK)	*Triangulating Instrumental Variable, confounder adjustment and Difference-in-Difference methods for comparative effectiveness research in observational data*
12:08-12:30	**IN5.4**	**Anthony Devaux** (FR)	*Random survival forests for competing causes with multivariate longitudinal endogenous covariates*

ROOM G.56

11:00-12:30 **PARALLEL SESSION 17: Competing risks**

Chair: **Jeremy Taylor** (US)

11:00-11:18	**S17.1**	**Alexandra Bühler** (CA)	*Estimands for Recurrent and Terminal Events: Methods, Issues and Recommendations*
11:19-11:36	**S17.2**	**Niklas Maltzahn** (NO)	*Separable Effects of Baseline Exposure in Multi-State Models*
11:37-11:54	**S17.3**	**Harry Parr** (UK)	*Externally Validating Clinical Dynamic Prediction Joint Models for Localised Prostate Cancer*
11:55-12:12	**S17.4**	**Damjan Manevski** (SI)	*Integrating relative survival in multi-state models – A non-parametric approach*
12:13-12:30	**S17.5**	**Andreas Gleiss** (AT)	*Degrees of necessity and of sufficiency for competing risks survival data*

ROOM G.06

11:00-12:30 **PARALLEL SESSION 18: Personalized Medicine**

Chair: **Ruth Keogh** (UK)

11:00-11:18	**S18.1**	**Ben Lanza** (UK)	*Resampling Methods to Control the Family Wise Error Rate for Dual Biomarker Threshold Identification*
11:19-11:36	**S18.2**	**Caterina Gregorio** (IT)	*Personalized optimal treatment timing through multi-state modelling and microsimulation*
11:37-11:54	**S18.3**	**Jennifer Hellier** (UK)	*Methods for estimating personalized treatment recommendations with extensions to survival data*
11:55-12:12	**S18.4**	**Luana Boumendil** (FR)	*Drugs combinations screening using a Bayesian ranking approach based on dose-response models*
12:13-12:30	**S18.5**	**Matias Janvin** (CH)	*Dynamic interventions determined by recurrent events*

ROOM 1.17

11:00-12:30 **PARALLEL SESSION 19: Clinical Trials**

Chair: **Alun Bedding** (UK)

11:00-11:18	**S19.1**	**Stella Jinran Zhan** (UK)	*Should the two-trial paradigm still be the gold standard in drug assessment?*
11:19-11:36	**S19.2**	**Charlotte Micheloud** (CH)	*Conditional Drug Approval with the Harmonic Mean Chi-Squared Test*
11:37-11:54	**S19.3**	**Richard Emsley** (UK)	*Frequentist and Bayesian approaches to rescuing disrupted trials*
11:55-12:12	**S19.4**	**Aaron Sarvet** (CH)	*The role of grace periods in comparative effectiveness studies of different medications*
12:13-12:30	**S19.5**	**Dawn Teare** (UK)	*Analysis of trials with intervention induced post randomisation clustering*

ROOM 2.16

11:00-12:30 **PARALLEL SESSION 20: Heterogeneity in effects**

Chair: **Philip Boonstra** (US)

11:00-11:18	**S20.1**	**Laura Savare** (IT)	*Evaluation of mental health patients' diagnostic-therapeutic paths through state sequences analysis*
11:19-11:36	**S20.2**	**Ashwini Venkatasubramaniam** (UK)	*Causal DART: A non-parametric Bayesian approach to estimate heterogeneous treatment effects*
11:37-11:54	**S20.3**	**Shaun Seaman** (UK)	*Adjusting for time of positive test when estimating the risk of a post-infection outcome*
11:55-12:12	**S20.4**	**Silvia Metelli** (FR)	*Disentangling interactions between components of complex health interventions*
12:13-12:30	**S20.5**	**Alexandros Rekkas** (NL)	*Predicting individualized treatment effects using baseline risk: A simulation study*

12:30-13:30 **LUNCH**

SOCIAL PROGRAMME

13:30-17:00 **SOCIAL EXCURSIONS**

- City walk around Newcastle Gateshead
- Victoria tunnels tour
- Hadrian's wall

ROOM G.41

09:00-10:30 **INVITED SESSION 6: Bayesian nonparametrics and ML for causal inference**

Chair: **Karla Diaz-Ordaz** (UK)

09:00-09:22	**IN6.1**	**Stijn Vansteelandt** (BE)	*Assumption-lean Cox regression*
09:23-09:45	**IN6.2**	**Alex Luedtke** (US)	*Adversarial Monte Carlo Meta-Learning of Conditional Average Treatment Effects*
09:46-10:07	**IN6.3**	**Jason Roy** (US)	*Bayesian nonparametric methods for causal inference*
10:08-10:30	**IN6.4**	**Samrachana Adhikari** (US)	*Nonparametric Bayesian Instrumental Variable Analysis: Evaluating Heterogeneous Effects of Coronary Arterial Access Site Strategies*

ROOM G.56

09:00-10:30 **PARALLEL SESSION 21: Communicating statistical concepts**

Chair: **John Matthews** (UK)

09:00-09:18	**S21.1**	**David Whitney** (UK)	*Why should I? Toward improved communication and evaluation of estimated dynamic treatment strategies*
09:19-09:36	**S21.2**	**Bernard Francq** (BE)	*P-value, s-value, b-value, d-value,... What else? Individual Success Probability*
09:37-09:54	**S21.3**	**Marissa LeBlanc** (NO)	*Statistical advising: professional development opportunities for the biostatistician*
09:55-10:12	**S21.4**	**Dominik Grathwohl** (CH)	*The inflation of p-values of likelihood-ratio tests in longitudinal data analysis*
10:13-10:30	**S21.5**	**Deepak Parashar** (UK)	*Visualising Master Protocols*

ROOM G.06

09:00-10:30 **PARALLEL SESSION 22: Counterfactuals**

Chair: **Camila Olarte Parra** (UK)

09:00-09:18	**S22.1**	**Ilaria Prosepe** (NL)	*Counterfactual simulation to evaluate sequential stratification methods*
09:19-09:36	**S22.2**	**Andrew Grant** (UK)	*Bias in multivariable Mendelian randomization studies due to measurement error on exposures*
09:37-09:54	**S22.3**	**Pawel Morzywolek** (BE)	*Sequential counterfactual prediction to support individualized decisions on treatment initiation*
09:55-10:12	**S22.4**	**Nan van Geloven** (NL)	*Assessing discrimination of counterfactual prediction models for time-to-event outcomes*
10:13-10:30	**S22.5**	**Ruth Keogh** (UK)	*Estimation and calibration of counterfactual risk predictions, with application to liver transplant*

ROOM 1.17

09:00-10:30 PARALLEL SESSION 23: Bias and Estimation

Chair: **Katherine Lee** (AU)

09:00-09:18	**S23.1**	**Heiko Goette** (DE)	*Estimation of treatment effects in randomized clinical trials involving external control data*
09:19-09:36	**S23.2**	**Jan Meis** (DE)	*Performance of different estimators in adaptive two-stage trials with optimized design parameters*
09:37-09:54	**S23.3**	**Monisha Dewan** (UK)	*Considerations for the design and analysis of nested case-control studies*
09:55-10:12	**S23.4**	**Ali Shariati** (AU)	*Confidence Band for the Cumulative Hazard Rate Function in Right Censored Length-biased Samplin*
10:13-10:30	**S23.5**	**Stina Zetterstrom** (SE)	*Selection bias and multiple inclusion criteria in observational studies*

ROOM 2.16

09:00-10:30 PARALLEL SESSION 24: Lightning never strikes twice

Chair: **Armin Schueler** (DE)

09:00-09:18	**S24.1**	**Wang Pok Lo** (UK)	*Surrogate endpoints in regulatory use: how many are actually statistically valid?*
09:19-09:36	**S24.2**	**Martin Gebel** (DE)	*Application of the R Shiny app DetectoR for signal detection and label generation*
09:37-09:54	**S24.3**	**Michael Lauseker** (DE)	*Internal pilot designs for sample size recalculations in animal experiments - prospects and pitfalls*
09:55-10:12	**S24.4**	**Steve Ferreira Guerra** (CA)	*Pragmatic SIMEX method to correct for measurement error in time-varying prescription-based exposures*
10:13-10:30	**S24.5**	**Emily Granger** (UK)	*Estimating the effects of multiple treatments in combination on outcomes using longitudinal data*

10:30-11:00 BREAK

ROOM G.41

11:00-12:00 KEYNOTE SPEAKER LECTURE: Using Electronic Health Records for Scientific Research: Promises and Perils

Introduction by: **Thomas Jaki** (UK), *ISCB43 SPC Chair*

Keynote Speaker: **Prof. Bhramar Mukherjee** (US), *University of Michigan School of Public Health*

12:00-13:30 LUNCH

ROOM G.06

12:15-13:15 ISCB ANNUAL GENERAL MEETING

ROOM G.41

13:30-15:00 INVITED SESSION 7: Semi-competing risks and causal inference

Chair: **Nan van Geloven** (NL)

13:30-13:55	**IN7.1**	**Mats Julius Stensrud** (CH)	*Estimands of practical interest in intercurrent event settings*
13:55-14:20	**IN7.2**	**Daniel O Scharfstein** (US)	*A Bayesian nonparametric approach for evaluating the causal effect of treatment in randomized trials with semi-competing risks*
14:20-14:45	**IN7.3**	**Linda Valeri** (US)	*A multistate approach for the study of interventions on an intermediate time-to-event in health disparities research*
14:45-15:00		**Els Goetghebeur** (BE)	**DISCUSSION**

ROOM G.56

13:30-15:00 PARALLEL SESSION 25: Simulation and software

Chair: **Daniel Sabanes Bove** (CH)

13:30-13:48	**S25.1**	**Elias Laurin Meyer** (AT)	*Simple – A modular, open-source R software solution to SIMulate PLatform trials Efficiently*
13:49-14:06	**S25.2**	**Rhys Bowden** (AU)	*Simulating binary variables with two levels of clustering*
14:07-14:24	**S25.3**	**Oliver Boix** (DE)	*A collaborative approach to software development: The crmPack experience*
14:25-14:42	**S25.4**	**Tom Parke** (UK)	*Simulation Guided Trial Design – The Challenges and The Benefits*
14:43-15:00	**S25.5**	**Aiden Smith** (UK)	*Improving data transparency in the research community by constructing synthetic time-to-event data*

WEDNESDAY 24 AUGUST 2022

ROOM G.06

13:30-15:00 PARALLEL SESSION 26: Missing data in Studies

Chair: **Susie Cro** (UK)

13:30-13:48	**S26.1**	**Mike Daniels** (US)	*A flexible approach for the analysis of repeated attempt designs*
13:49-14:06	**S26.2**	**Mia Tackney** (UK)	*Non-parametric Multiple Imputation for Epoch-level Wearable data in Trials*
14:07-14:24	**S26.3**	**Camila Olarte Parra** (UK)	*Targeting hypothetical estimands with causal inference and missing data estimators in a real trial*
14:25-14:42	**S26.4**	**Melissa Middleton** (AU)	*Evaluation of multiple imputation approaches for case-cohort studies with binary outcomes*
14:43-15:00	**S26.5**	**Audinga-Dea Hazewinkel** (UK)	*Outcome variances after dropout as an indicator of dropout bias in randomized controlled trials*

ROOM 1.17

13:30-15:00 PARALLEL SESSION 27: Prediction modelling

Chair: **Ewout Steyerberg** (NL)

13:30-13:48	**S27.1**	**Doranne Thomassen** (NL)	*Effective sample size: a measure of individual uncertainty in predictions*
13:49-14:06	**S27.2**	**Lauren Coan** (UK)	*Automatic detection of glaucomatous neuropathy from fundus images*
14:07-14:24	**S27.3**	**Philip Boonstra** (US)	*A comparison of methods for incorporating information from historical prediction models*
14:25-14:42	**S27.4**	**Oliver Church** (UK)	*Comparing methods to generate disease risk factor trajectories for longitudinal microsimulations*
14:43-15:00	**S27.5**	**Leiv Rønneberg** (UK)	*Dose-response prediction for in-vitro drug combination datasets: a probabilistic approach*

ROOM 2.16

13:30-15:00 PARALLEL SESSION 28: Efficient trial designs

Chair: **James Wason** (UK)

13:30-13:48	**S28.1**	**Kazufumi Okada** (JP)	*Two one-sided test-then-pool method for clinical trials*
13:49-14:06	**S28.2**	**Michael LeBlanc** (US)	*Using the Probability of Longer Survival to Assess the Efficacy of New Cancer Therapies*
14:07-14:24	**S28.3**	**Thomas Jaki** (UK)	*Testing for treatment differences with allocation probabilities in response adaptive trials*
14:25-14:42	**S28.4**	**Becky Turner** (UK)	*Determining sample size in a personalised randomised trial comparing a network of treatments*
14:43-15:00	**S28.5**	**Jeremy Taylor** (US)	*Solutions for Surrogacy Validation with Longitudinal Outcomes for a Gene Therapy*

15:00-15:30 BREAK

ROOM G.41

15:30-17:00 PARALLEL SESSION 29: Joint inference with high dimensional data

Chair: **Simon White** (UK)

15:30-15:48	**S29.1**	**F. Javier Rubio** (UK)	*Bayesian variable selection with applications to selection of comorbidities*
15:49-16:06	**S29.2**	**Camilla Lingjaerde** (UK)	*A scalable ECM algorithm for multiple-network joint inference with the graphical horseshoe*
16:07-16:24	**S29.3**	**Alessandro Gasparini** (SE)	*A Natural History and Copula Based Joint Model for Regional and Distant Breast Cancer Metastasis*
16:25-16:42	**S29.4**	**Alexandra Lavalley-Morelle** (FR)	*Marker selection in joint analysis with competing risks: application to SARS-COV-2 patients*
16:43-17:00	**S29.5**	**Roula Tsonaka** (NL)	*Modelling dynamic associations in multivariate longitudinal data*

ROOM G.56

15:30-17:00 PARALLEL SESSION 30: Machine learning and prediction

Chair: **Manuela Zucknick** (NO)

15:30-15:48	**S30.1**	**Francesca Ieva** (IT)	*Deep Survival EWAS modeling of cancer time to diagnosis via blood-derived DNA methylation profiles*
15:49-16:06	**S30.2**	**Francois Grolleau** (FR)	*A Comprehensive Framework for the Evaluation of Individual Treatment Rules from Observational Data*
16:07-16:24	**S30.3**	**Richard Riley** (UK)	*Quantifying instability after developing a clinical prediction model*
16:25-16:42	**S30.4**	**Awa Diop** (CA)	*History-Restricted MSM and LCGM of Treatment Trajectories for a time-dependent outcome*
16:43-17:00	**S30.5**	**Mariella Gregorich** (AT)	*Flexible parametrization of individual sparsified networks for prediction: a proof-of-concept*

ROOM G.06

15:30-17:00 PARALLEL SESSION 31: Early Phase Trials

Chair: **Matthew Schipper** (US)

15:30-15:48	**S31.1**	**Andrew ROOM** (UK)	*Decision making under uncertainty in PI-II dose finding trials in Oncology*
15:49-16:06	**S31.2**	**Helene Thygesen** (UK)	*Beyond 3+3 – acceptability and implementation of model-based dose-finding study designs in practice*
16:07-16:24	**S31.3**	**Cornelia Ursula Kunz** (DE)	*Advanced tumor metrics to support characterization of the dose-response relationship*
16:25-16:42	**S31.4**	**Dario Zocholl** (DE)	*Improving interim decision in two-stage phase II designs by incorporating short-term endpoints*
16:43-17:00	**S31.5**	**Nina Wilson (UK)**	*Aggregating prior distributions from experts for sample size calculations*

ROOM 1.17

15:30-17:00 PARALLEL SESSION 32: Complex Modelling

Chair: **Elisavet Syriopoulou** (SE)

15:30-15:48	**S32.1**	**Freya Tyrer** (UK)	*Flexible parametric methods for calculating life expectancy in small populations*
15:49-16:06	**S32.2**	**Kishor Das** (IE)	*Functional Limits of Agreement using a Mixed Effects Modelling Framework in Method Comparison Study*
16:07-16:24	**S32.3**	**Benoît Sauty** (FR)	*Progression models for imaging data with Longitudinal Variational Auto Encoders*
16:25-16:42	**S32.4**	**Annah Muli** (UK)	*Modelling the effect of time varying covariates in time to event studies of twins with delayed entry*
16:43-17:00	**S32.5**	**Kim Luijken** (NL)	*Quantitative prediction error analysis to investigate performance under measurement heterogeneity*

CIVIC CENTRE NEWCASTLE

19:00-23:00 CONFERENCE DINNER

ONLINE

19:00-19:45 History Talk: Meet the story of Coal to Culture for online attendees

ROOM G.06

09:00-12:30 MINI SYMPOSIUM 1: STRengthening Analytical Thinking for Observational Studies (STRATOS) initiative

Organizers: **Willi Sauerbrei**, *Freiburg, DE*
Ruth Keogh, *London, UK*

09:00-09:05	**Ruth Keogh** (UK)	*Short introduction and overview of the program*

MS1

09:05-09:30	**Lara Lusa** (SI) **Kate Lee** (AU) **Carsten Schmidt** (DE) **Marianne Huebner** (US)	*Check what is missing with initial data analysis*

Topic Group: "Initial Data Analysis" (TG3)

09:30-09:55	**Roderick Little** (US) **James Carpenter** (UK) **Katherine Lee** (AU)	*A comparison of three popular methods for handling missing data: complete-case analysis, inverse probability weighting, and multiple imputation*

Topic Group: "Missing data" (TG1)

09:55-10:20	**Victor Kipnis** (US) **Laurence Freedman** (IL & US) **Pamela Shaw** (US)	*Berkson measurement error and the regression calibration approach*

Topic Group: "Measurement Error and Misclassification" (TG4)

10:20-10:45	**Aris Perperoglou** (UK) **Daniela Dunkler** (AT) **Christine Wallisch** (AT) **Matthias Schmid** (DE) **Patrick Royston** (UK) **Willi Sauerbrei** (DE)	*Comparison of Multivariable Fractional Polynomials with Splines and Penalised Spines*

10:45-11:00 BREAK

Topic Group: "Selection of variables and functional forms in multivariable analysis" (TG2)

11:00-11:25	**Nan van Geloven** (NL)	*Validation of prediction models in the presence of competing risks*

A collaborative project of Topic Groups "Evaluating diagnostic tests and prediction models" (TG6) and "Survival analysis" (TG8)

11:25-11:50	**Michal Abrahamowicz** (CA) **Marie-Eve Beauchamp** (CA) **Anne-Laure Boulesteix** (DE) **Tim P. Morris** (UK) **Willi Sauerbrei** (DE) **Jay S. Kaufman** (CA)	*Use of data-driven simulations to inform real-world survival analyses*

Panel on "Simulation Studies"

11:50-12:15	**Saskia le Cessie** (NL) **Els Goetghebeur** (BE) **Limin Liu** (BE) **Doranne Thomassen** (NL) on behalf of **work package 3 of the Setting International Standards in Analyzing Patient-Reported Outcomes and Quality of Life Endpoints Data Consortium (SISAQOL)**	*Developing international standards in the analysis of patient reported outcomes in cancer clinical trials: methodological issues and STRATOS engagement in the European IMI-SISAQOL project*

Discussion about the next steps

12:15-12:30	**Willi Sauerbrei** (DE)	**DISCUSSION**

ROOM G.41

09:00-12:30 MINI SYMPOSIUM 2: Modern software tools for modern statistics

Organizers: **Michael Crowther**, Stockholm, SE
Svetlana Cherlin, Newcastle, UK
Thomas Jaki, Cambridge, UK

MS2

09:00-09:30	**Duco Veen** (NL)	*Why I use ... Stan*
09:30-10:00	**Maren Hackenberg** (DE)	*Why I use ... Julia*
10:00-10:30	**Paul Zivich** (US)	*Why I use ... Python*
10:30-11:00		**BREAK**
11:00-11:45	**Kevin Kunzmann** (DE)	*Github: Reproducibility and beyond*
11:45-12:30	**Yulia Marchenko**	*Professional software development: Tips and tricks*

ROOM G.56

09:00-12:30 ECB DAY

PART A

Chair: **TBA**

09:00-09:45	**ECB1** ***Invited***	**John Matthews** (UK)	*Biostatistics: always changing, always staying the same*
09:45-10:00	**ECB2** ***OC***	**Heather Poad** (UK)	*Experiences from a systematic review of pharmacological therapies in colorectal cancer*
10:00-10:15	**ECB3** ***OC***	**Alessandro Gasparini** (SE)	*Getting Comfortable with Being Uncomfortable... as an Early Career Biostatistician*
10:15-10:30	**ECB4** ***OC***	**Jeroen Mulder** (NL)	*Navigating statistical methods across different disciplines*
10:30-11:00			**BREAK**

PART B

Chair: **TBA**

11:00-11:15	**ECB5** ***OC***	**Myra McGuinness** (AU)	*Peer Review Process for Manuscripts*
11:15-11:30	**ECB6** ***OC***	**Judith ter Schure** (NL)	*What can biostatisticians do to reduce research waste?*
11:30-12:15	**ECB7** ***Invited***	**Noah Haber** (US)	*Causation and collaboration: making of a 48-person, 1000+ article, systematic interdisciplinary evaluation exploring the state of causal language in health research*
12:15-12:30			PANEL DISCUSSION / Q&A

President's Invited Speaker

On the many uses and abuses of regression models

John Carlin[1]

[1] *Professorial Fellow, Murdoch Children's Research Institute & The University of Melbourne, Australia*

As a PhD student, I was puzzled when my supervisor told me that techniques are not the important things in statistics. What he meant is that a statistician who is collaborating with an empirical scientist should focus on gaining a deep understanding of the substantive problem that the collaborator is seeking to solve, rather than turning immediately to the bank of technical knowledge about models and methods that s/he might have acquired. Our practice and teaching of statistics often fail to recognise the essential truth of this point. Recently this has begun to change, and the key issues are articulated particularly well in Hernan's elucidation of the "three tasks of data science" (description, prediction, causal inference). However, we have a long way to go, evidenced by the fact that the applied literature is still full of papers that fit multivariable regression models for ill-defined purposes such as "identifying risk factors" or "exploring the effects" of a list of variables on an outcome. This talk will focus on the deceptively familiar (generalised) regression models that dominate statistical practice, first asking what exactly they are and then considering how and why they might be used to address each of the three types of empirical question. Is our aim to identify the correct model, or is it to develop a model that is useful for a particular purpose? How should that influence our approach to model specification and interpretation? I will then look at how we teach these fundamental concepts to students (and to colleagues) and how we might reform that teaching, in order to encourage better practice and thus better science..

Keynote Invited Speaker

Using Electronic Health Records for Scientific Research: Promises and Perils

Bhramar Mukherjee[1]

[1] *John D Kalbfleisch Collegiate Professor of Biostatistics, School of Public Health, University of Michigan, Ann Arbor MI, United States*

Electronic Health Records (EHR) linked with other auxiliary data sources hold tremendous potential for conducting real time actionable research. However, one has to answer two fundamental questions before conducting inference: "Who is in my study?" and "What is the target population of Inference?". Without accounting for selection bias, one can quickly produce rapid but inaccurate conclusions. In this talk, I will discuss a statistical framework for jointly considering selection bias and phenotype misclassification in analyzing EHR data. Examples will include genome and phenome-wide association studies of Cancer and COVID-19 outcomes using data from the Michigan Genomics initiative and the UK Biobank. This is joint work with Lars Fritsche, Lauren Beesley and Maxwell Salvatore at the University of Michigan School of Public Health.

Invited Sessions

INVITED SESSION 1: Machine learning with small data

IN1.1 Structured priors for improving treatment response prediction in cancer pharmacogenomic screens

Manuela Zucknick[1], Zhi Zhao[1,2], Marco Banterle[3], Alex Lewin[3]

[1] *University of Oslo, Norway,* [2] *Oslo University Hospital, Norway,* [3] *London School of Hygiene and Tropical Medicine, United Kingdom*

Large-scale cancer pharmacogenomic screening experiments profile cancer cell lines or patient-derived cells versus hundreds of individual drug compounds or drug combinations. The aim of these in vitro studies is to use the genomic profiles of the cell samples together with information about the drugs to predict the response to a particular treatment, for example to identify combinations of drugs that act synergistically. Unfortunately, the in vitro cell viability experiments that are performed to measure drug response are technically challenging and expensive, and even the largest drug screens only include a few dozen or at most hundreds of cell samples. This means that we are in a multi-response regression setting with few samples and high-dimensional heterogeneous (multi-omics) input data.
In this challenging setting we aim to improve model performance by using information about structure in the data, through designing structured penalties or Bayesian priors for combining the different genomic data sources efficiently in the input matrix, for borrowing information across correlated drug response variables, and for including external knowledge about biological structures such as drug target pathways. In this talk I will explore a multivariate Bayesian variable and covariance selection modelling setup to achieve this aim, for which we have provided an efficient implementation as an R package (https://CRAN.R-project.org/package=BayesSUR). In particular, we propose to make use of known structure between responses and predictors, e.g. molecular pathways and related omics features targeted by specific drugs, via a Markov random field prior for the latent variable selection indicators of the coefficient matrix in sparse seemingly unrelated regression.

IN1.2 Sequential learning of regression models through penalized estimation

Wessel van Wieringen[1], Harald Binder[1]

[1] *Amsterdam UMC, Netherlands*

Online regression learning is concerned with the updating of current estimate of the generalized linear model's regression parameter when novel data become available. In particular, we consider regression modeling settings where parameter estimates from previous data can be used as anchoring points, yet may not be available for all parameters, thus covariance information cannot be reused. Hereto we present a procedure that updates through targeted penalized estimation, which shrinks the estimator towards a nonzero value. The current parameter estimate serves as this nonzero shrinkage target when an update is sought from novel data. This updating procedure can incorporate other current estimates of the parameter produced by competing (online) regression learning procedures. The iteratively updated regression parameter estimator is shown to be asymptotically unbiased and consistent. The penalty parameter is chosen through constrained cross-validated loglikelihood optimization. The constraint bounds the amount of shrinkage of the updated estimator toward the current one from below, which aims to preserve the (updated) estimator's goodness-of-fit on all-but-the-novel data. The proposed approach is compared both in papyro and in silico to other regression modeling procedures. Finally, we illustrate the proposed procedure on the large epidemiological fingertips study, which registers annual information on a range of public health indicators for the counties of England (see https://fingertips.phe.org.uk/). The annually arriving novel data batches have a different covariate-availability, which is accommodated by our procedure. The proposed procedure has been implemented and made available through the R-package porridge.

IN1.3 Interpretable deep generative approaches for small omics data-sets

Moritz Hess[1], Jens Nußberger[1], Frederic Boesel[1], Maren Hackenberg[1], Stefan Lenz[1], Harald Binder[1]

[1] *Institute of Medical Biometry and Statistics Faculty of Medicine and Medical Center, University of Freiburg, Germany*

High dimensional omics data are now frequently generated in the clinical routine, e.g., for tumor phenotyping as well as in biomedical research projects, e.g., to better understand an experimental setting on the level of single cells. Although the high dimensionality of omics data allows for a detailed characterization, e.g., of gene activity, the data are hard to interpret for the same reason.
Deep generative models (DGMs) can unsupervised learn latent variables from the high number of variables observed in omics data, resulting in a low dimensional, better interpretable representation composed of few latent variables. In order to link this representation, inferred from the data, with domain-expert knowledge and the literature, we need to identify the relationships between the inferred latent and the observed variables.
However, as the mapping from the observed to the latent space learnt by the DGMs is non-linear, identifying relationships between latent and observed variables is challenging.
To address these challenges, we here demonstrate, based on single-cell gene expression data, how latent variables learnt by deep generative models can be annotated in terms of the observed variables. Specifically, we investigate the joint distribution of observed and latent variables with log-linear models and perform variable selection in a step-wise manner. This allows us to identify multivariable gene patterns that are related to the latent variables.
As fitting DGMs to omics data is challenging, in particular due to the still rather low number of available observations caused by data privacy restrictions or high experimental costs, we further show, how well popular deep generative approaches, specifically deep Boltzmann machines (DBMs), variational autoencoders (VAEs) and generative adversarial networks (GANs) learn the structure of omics data given sample size constraints.

INVITED SESSION 2: Master protocols

IN2.1 Statistical issues in design and analysis of master protocols

James Wason[1]

[1] *Newcastle University, United Kingdom*

New statistical methods and trial designs have a big role to play in improving clinical trials. They can lead to more efficiency, better evidence and better outcomes of patients enrolled on the trial. This has been highlighted by the COVID-19 pandemic, where innovative methods such as adaptive platform trials have played a major role in improving treatment.
In this talk, I will provide a short overview of an innovative class of trial design called master protocols. Master protocols allows combining several separate but related clinical trials together. This provides substantial operational and statistical advantages. However, it also introduces several statistical issues that new methods are required for addressing. Some of these issues will be considered in more detail by subsequent speakers in the session.

Invited Sessions

IN2.2 Multiplicity Issues in Exploratory Platform Trials – Online Control of the False Discovery Rate

Franz Koenig[1], Sonja Zehetmayer[1], Martin Posch[1]

[1] *Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Austria*

When testing multiple hypotheses, a suitable error rate should be controlled even in exploratory trials. Conventional methods to control the False Discovery Rate (FDR) assume that all p-values are available at the time point of test decision. In platform trials, however, treatment arms enter and leave the trial at different times during its conduct. Therefore, the actual number of treatments and hypothesis tests is not fixed in advance and hypotheses are not tested at once, but sequentially. Recently, for such a setting the concept of online control of the FDR was introduced.
We propose several heuristic variations of the LOND procedure (significance Levels based On Number of Discoveries) that incorporate interim analyses for platform trials, and study their online FDR via simulations. To adjust for the interim looks spending functions are applied with O'Brien-Fleming or Pocock type group-sequential boundaries. The power depends on the prior distribution of effect sizes, e.g., whether true alternatives are uniformly distributed over time or not. We consider the choice of design parameters for the LOND procedure to maximize the overall power and investigate the impact on the FDR by including both concurrent and non-concurrent control data.

IN2.3 Bayesian Basket Trial Design with False Discovery Rate Control

Emily Zabor[1], Michael J Kane[2], Satrajit Roychoudhury[3], Lei Nie[4], Brian P Hobbs[5]

[1] *Cleveland Clinic, Cleveland OH, United States,* [2] *Yale University, New Haven CT, United States,* [3] *Statistical Research and Data Science Center, Pfizer Inc., New York NY, United States,* [4] *U.S. Food and Drug Administration, Silver Spring MD, United States,* [5] *Dell Medical School, The University of Texas at Austin, Austin TX, United States*

Oncology therapies have evolved over time, from cocktails of cytotoxic drugs to non-cytotoxic therapies that target specific pathways in tumor cells or promote anti-cancer immunity. Drug developers endeavor to understand treatment benefit heterogeneity at the patient level and establish actionable targets through biomarker-guided therapies. Recent advances in developing "tumor agnostic" treatments have identified molecular targets that define patient subpopulations in a manner that supersedes conventional criteria for cancer classification. These successes have produced effective targeted therapies that are administered to patients regardless of their tumor histology. Master protocol designs have been used to design more efficient trials of many biomarker-targeted therapies. Basket trials are one type of master protocol design where a single treatment is tested across multiple "baskets" of patients, where the baskets are often defined by different cancer histologies. By blending translational and clinical science, basket trials are well-suited to the task. But the statistical methods applied to basket trials have often treated the baskets as independent, an assumption that is unlikely in many settings. In addition, the multiple testing that is inherent to basket trial design has often been ignored. In this talk I will introduce guidance for the design and analysis calibration of basket trials based on the multisource exchangeability model. I will also present an approach to design basket trials that control the false-discovery rate. The methodology is applied to a case study based on the SUMMIT basket trial.

IN2.4 On the Lung Cancer Master Protocol (Lung-MAP)

Mary Redman[1]

[1] *Fred Hutchinson Cancer Center, Seattle WA, United States*

The Lung Cancer Master Protocol (Lung-MAP), initiated in June of 2014, was the first master protocol implemented within the United States National Cancer Institute (NCI) Network. Lung-MAP is an umbrella, biomarker-driven platform master protocol conducted through a public-private partnership involving the SWOG Cancer Research Network, the Alliance, ECOG-ACRIN, NRG Oncology, NCI, Foundation for the NIH, and Friends of Cancer Research. It was originally designed to evaluate investigational treatments for biomarker-defined populations in 2nd line stage IV or recurrent squamous cell lung cancers (SCC) within independently conducted and analyzed sub-studies. The sub-studies used a common phase II/III design for initial screening for activity and potential seamless definitive evaluation with the potential for regulatory approval for successful biomarker/biomarker-driven treatment pairs.
Beginning with the approval of immunotherapy for previously-treated lung cancers in 2015, Lung-MAP has undergone a series of changes to remain relevant to the ever-changing landscape of treatment for non-small cell lung cancer (NSCLC). The modifications to the study are numerous. The study expanded from 2nd line only to all previously treated squamous cell lung cancers, to all histologic types of NSCLC from SCC. The study shifted from a focus on seamless progression to definitive evaluation to more staged screening and potentially subsequent definitive evaluation. The study expanded from evaluating immunotherapy in immunotherapy naïve patients to studies designed to overcome resistance to immunotherapy. Finally, the study expanded to focus on combinations of treatments for both biomarker-driven populations and for immunotherapy-relapsed patients.
Within the Lung-MAP infrastructure, 16 studies have been initiated, 9 have been completed and published, an additional 2 presented at a national meeting, 2 were closed early, and 3 are currently accruing patients. Approximately 4500 patients have undergone genomic screening as part of Lung-Map and 1025 have registered to a treatment sub-study. We presented the findings of a positive randomized phase II evaluating an immunotherapy combination at ASCO 2022 and are planning a follow-on phase III.
The complex infrastructure of Lung-MAP requires active involvement by our scientific leadership (clinical, statistical, and translational medicine), our drug selection committee, site coordinators, and many other members of the study team. But importantly, this infrastructure also enables the evaluation of therapies in rare population and efficient implementation of new ideas..

INVITED SESSION 3: Best practices in statistical simulation and computing

IN3.1 On the researchers‘ degree of freedom in comparison studies

Anne-Laure Boulesteix[1], Christina Nießl[1], Theresa Ullmann[1], Simone Furcas[1]

[1] *LMU Munich, Germany*

In recent years, the need for neutral benchmark studies that focus on the comparison of methods coming from computational sciences has been increasingly recognized by the scientific community. While general advice on the design and analysis of neutral benchmark studies can be found in recent literature, a certain flexibility always exists. This includes the choice of data sets and performance measures, the handling of missing performance values, and the way the performance values are aggregated over the data sets. As a consequence of this flexibility, researchers may be concerned about how their choices affect the results or, in the worst case, may be tempted to engage in questionable research practices (e.g., the selective reporting of results or the post hoc modification of design or analysis components) to fit their expectations. We use two example benchmark studies to illustrate 1) how variable benchmark results can be when all possible combinations of a range of design and analysis options are considered; 2) how simulations can be simply manipulated in order to claim the superiory of a favorite (new) method. A particular emphasis will be placed on the handling of missing values in the benchmarking results arising when methods do not yield valid results for all considered datasets or simulation repetitions.

Invited Sessions

IN3.2 Neutral schmeutral: fair comparisons and simulation-study estimands

Tim P Morris[1], Brennan C. Kahan[1], Tra My Pham[1]

[1] *MRC Clinical Trials Unit at UCL, United Kingdom*

For simulation studies that aim to provide a fair comparison of methods, researchers must make a number of decisions about the specific implementation of each method. The performance of the method may depend on these decisions. These researchers are unlikely to be equally expert in all methods and may have more knowledge about the methods than those who would use them in practice. This leads to a dilemma about the comparison of interest. In running a simulation study, should we implement methods as an expert or using ‘factory settings’ to evaluate the performance of a procedure that analysts might use?
We highlight these issues using a published simulation study. The authors considered an interesting broad question about ‘statistical’ vs. ‘machine learning’ methods for missing data imputation but made some dubious choices in implementing the methods. We consider these choices by drawing a parallel between interventions compared in clinical trials and methods compared in simulation studies. By doing so, we can think about the estimand of a simulation study. We work through five attributes parallel to those defined in the ICH E9(R1) addendum. This structured thought about a simulation study’s estimand takes focus away from the goal of a completely ‘fair’ or ‘neutral’ comparison, which will always be questionable. It instead forces us to be clear about choices and, importantly, their rationale when designing a simulation study.

IN3.3 Step-by-step guidance on best practices in statistical computing

Beth Ann Griffin[1], Ricardo Sanchez[2], Daniel McCaffrey[3], Joseph Pane[1]

[1] *RAND, Santa Monica CA, United States,* [2] *UnitedHealth Group,* [3] *Educational testing services*

The world is becoming increasingly complex, both in terms of the rich sources of data we have access to and the statistical and computational methods we can use on data. These factors create an ever-increasing risk for errors in code and the sensitivity of findings to data preparation and the execution of complex statistical and computing methods. The consequences of coding and data mistakes can be substantial. In this talk, we describe the key steps for implementing a code quality assurance (QA) process that researchers can follow to improve their coding practices throughout a project to assure the quality of the final data, code, analyses, and results. These steps include: (i) adherence to principles for code writing and style that follow best practices; (ii) clear written documentation that describes code, workflow, and key analytic decisions; (iii) careful version control; (iv) good data management; and (v) regular testing and review. Following these steps will greatly improve the ability of a study to assure results are accurate and reproducible. The responsibility for code QA falls on individual researchers, institutions, journals, and funding agencies.

Invited Sessions

INVITED SESSION 4: Ageing

IN4.1 Translating models of ageing into routine clinical practice

Andrew Clegg[1]

[1] *University of Leeds, United Kingdom*

Models of ageing have historically been problematic to translate into routine clinical practice as they have typically required additional clinical information to be collected as part of routine care, which is challenging given the time constraints of clinical practice. A novel approach has been to operationalise the cumulative deficit model of frailty as a model of ageing using routine electronic health record (EHR) data. In this presentation, Prof Clegg will outline the development, validation and national implementation of the electronic frailty index (eFI) using routine EHR data from around 1M UK patients. The national availability of the eFI provided the infrastructure for a major health policy change in the UK, with routine identification and management of older people with frailty embedded in national health policy as an international exemplar. Prof Clegg will also outline the statistical methods being used for the development of eFI2, which will address the limitations of the original eFI.

IN4.2 A new statistical framework for biological age estimation

Mar Rodríguez-Girondo[1], Marije Sluiskes[1], Hein Putter[1], Jelle Goeman[1]

[1] *Leiden University Medical Center, Netherlands*

Among people of the same chronological age, there is high variability in the rate of ageing. Whereas some individuals enjoy long and healthy lives, others suffer from diseases and die prematurely. The concept of biological age is postulated to better capture this variability, and hence to be a better indicator of an individual's true ageing status than chronological age and a better predictor of mortality and age-related diseases. Identification of biomarkers of biological age is a central topic in current aging research. However, since biological age is an elusive concept given its latent nature, it is unclear how to operationalize it (i.e. define an appropriate estimand) and systematically study the methodological properties of candidate methods for its prediction.
In this talk, I will present a new methodological framework for the conceptualization and analysis of biological age data, incorporating advanced survival analysis techniques into the omics-based prediction modeling framework. Our hypothesis is that biological age, of latent and holistic nature, can be operationalized in terms of observable time-to-event outcomes such as age-at-onset profiles of age-related diseases and/or mortality. I will illustrate this idea using simulated and real metabolomics data and I will compare our new methods to related approaches in the aging literature, namely the so-called "second generation" epigenetic clocks.

IN4.3 From variant to function: Identification and characterisation of genetic variants linked to human longevity

Joris Deelen[1]

[1] *Max Planck Institute for Biology of Ageing, Cologne, Germany*

Advancing age is the major risk factor for many serious illnesses, including cardiovascular disease and dementia. However, individuals that reach an exceptional old age often seem to escape or delay age-related diseases, and part of this trait seems to be encoded in their genome. We, and others, previously performed several large genome-wide association studies (GWAS) to identify genetic variants involved in longevity and related traits. However, the common variants identified through GWAS only explain a minor fraction of the heritability of longevity. The field is therefore switching towards identification of rare genetic variants in exceptionally long-lived individuals or human families exhibiting longevity. These studies mostly focus on genetic variants that are shared between long-lived family members (identified through, for example, linkage analysis) or on variants in evolutionary conserved lifespan-associated pathways, such as insulin/insulin-like growth factor 1 signalling. However, given the low frequency of rare variants it is often not possible to determine their causality in humans and functional studies in cellular models and, subsequently, model organisms are indispensable. In my group, we apply CRISPR/Cas9 gene editing to generate transgenic cell lines and mice that harbour genetic variants identified in long-lived individuals and we are currently investigating their effects on general health and lifespan.

INVITED SESSION 5: Student Awardees

IN5.1 Spatial Dependence Modeling of Latent Susceptibility and Time to Joint Damage in Psoriatic Arthritis

Fangya Mao[1], Richard Cook[1]

[1] *University of Waterloo, Waterloo ON, Canada*

Important scientific insights into chronic diseases affecting several organ systems can be gained from modeling spatial dependence of sites experiencing damage progression. We describe models and methods for studying spatial dependence of joint damage in psoriatic arthritis (PsA). Since a large number of joints may remain unaffected even among individuals with a long disease history, spatial dependence is first modelled in latent joint-specific indicators of susceptibility. Among susceptible joints, a Gaussian copula is adopted for dependence modeling of times to damage. Composite likelihoods are developed for settings where individuals are under intermittent observation and progression times are subject to type K interval censoring. Two-stage estimation procedures help mitigate the computational burden arising when a large number of processes (i.e. joints) are under consideration. Simulation studies confirm that the proposed methods provide valid inference, and an application to the motivating data from the University of Toronto Psoriatic Arthritis Clinic yields important which can help physicians distinguish PsA from arthritic conditions with different dependence patterns.

IN5.2 Sharing information across patient subgroups to draw conclusions from sparse treatment networks

Theodoros Evrenoglou[1], Silvia Metelli[1], Johannes Schneider-Thoma[2], Spyridon Siafis[2], Stefan Leucht[2], Anna Chaimani[1]

[1] Inserm Research Center of Epidemiology and Statistics, Université de Paris, France, [2] Department of Psychiatry and Psychotherapy, School of Medicine, Technical University of Munich, Germany

Network meta-analysis (NMA) usually provides estimates of the relative effects with the highest possible precision. However, poorly connected networks with limited data that fail to provide useful conclusions may arise under certain situations. This is often the case for 'sensitive' subgroups of the population that cannot be easily included in trials, such as children, elder patients, or individuals with multi-morbidity. Results from such networks are accompanied with substantial uncertainty not only in the estimation of the relative effects but also in the plausibility of the underlying assumptions since their formal evaluation is impossible. In addition, the large sample approximations on which NMA models typically rely on fail in the presence of only a handful of studies per comparison. Hence, lack of robustness and potentially limited reliability are common issues that might be encountered when analyzing sparse treatment networks. We propose a Bayesian framework that allows to share information between two networks that pertain to different population subgroups. Specifically, we use the results from a subgroup with a lot of direct evidence forming a 'dense' network to construct informative priors for the relative effects of the target subgroup forming a sparse network. This is a two-stage approach where at the first stage we synthesize the dense network using a hierarchical NMA model, in which we add a location parameter that shifts the distribution of the relative effects to make them applicable to the target population. At the second stage, these results are used as prior information for the sparse network. We inform the location parameter either through the data or using expert opinion. We illustrate our approach through an example of psychiatric patients where the target subgroup of children-adolescents forms a sparse network and the subgroup of "general patients" forms a dense network. Our approach results in much more precise and robust estimates of the relative effects which can give insight about the efficacy of antipsychotics for the target subgroup and adequately inform clinical practice.

IN5.3 Triangulating Instrumental Variable, confounder adjustment and Difference-in-Difference methods for comparative effectiveness research in observational data

Laura Guedemann[1], John Dennis[1], Andrew McGovern[1], Lauren Rodgers[1], Beverley Shields[1], Jack Bowden[1]

[1] University of Exeter, United Kingdom

In a typical clinical trial, the randomization of participants to different pharmacological treatment arms generally leaves them well-balanced with respect to possible confounders, which makes their analysis and interpretation straightforward. However, when assessing the effectiveness of competing treatments in observational data, confounding is omnipresent, potentially severe and cannot be simply ignored. Direct confounder adjustment via multivariable regression and Propensity score methods are the most popular option to address this. Instrumental Variable (IV) methods, leveraging for example a physician's preference for one treatment over another, and Difference-in-Difference (DiD) approaches, which leverage outcome data pre- and post-treatment, both circumvent the need for direct confounder adjustment altogether. Each method relies on a different set of assumptions and usually exploit different parts of the data, which makes them complimentary tools for causal triangulation.
We assess the impact of violating important assumptions of each method within a comprehensive simulation study. Furthermore, we propose the prior outcome augmented Instrumental Variable (POA-IV), a method that leverages physician preference IV data as well as data pre-and post- treatment, which is robust to the violation of the standard IV and DiD assumptions (IV is directly related to outcome and prior outcome influences treatment decision). We also propose the use of a generalised heterogeneity statistic to decide if two or more estimates are statistically similar, taking into account their correlation, in order to provide a quantitative basis for triangulating their findings. We illustrate our framework with an application to observational data from the UK Clinical Practice Research Datalink, to assess the absolute risk of genital infection side-effects in patients initiating Sodium-glucose Co-transporter-2 inhibitors (n = 2,487) versus Dipeptidyl Peptidase-4 inhibitors (n = 4,511) as second-line treatment for type 2 diabetes. Our novel triangulation-based approach has great potential to improve the robustness of comparative effectiveness studies of medications using large-scale observational health data.

IN5.4 Random survival forests for competing causes with multivariate longitudinal endogenous covariates

Anthony Devaux[1], Cécile Proust-Lima[1], Robin Genuer[2]

[1] Inserm BPH U1219, Université de Bordeaux, France
[2] Inserm U1219, Bordeaux Population Health, France

The individual data collected throughout patient follow-up constitute crucial information for assessing the risk of a clinical event, and eventually for adapting a therapeutic strategy. Joint models have been proposed to compute individual dynamic predictions from repeated measures to one or two markers. However, they hardly extend to the case where the complete patient history includes much more repeated markers. Our objective was thus to propose a solution for the dynamic prediction of a health event that may exploit repeated measures of a possibly large number of endogenous markers.
We extended the random survival forest methodology to incorporate multivariate longitudinal endogenous markers. The random survival forest is composed by an ensemble of decision trees, where the subjects are recursively split into two subgroups depending on the predictor that maximizes a split criterion, here the Fine and Gray statistic test to handle competing events. At each split, mixed models for the longitudinal markers are fitted and the predicted random effects are used among the others time-fixed predictors to split the subjects. The individual-specific event prediction is derived as the average over all trees of the leaf-specific cumulative incidence function computed using the Aalen-Johansen estimator. Predictive performance is assessed by cross-validation using estimators of Brier Score and the area under the Receiver Operating Characteristic curve adapted to censored data.
We demonstrate in a simulation study the performances of our methodology, both in a small dimensional context in comparison with classical joint models, and in a large dimensional context in comparison with two stage regression calibration approach. The method is applied to predict the individual risk of dementia in the elderly (accounting for the competing death) according to the trajectories of cognitive functions, brain imaging markers, and general clinical evaluation. Our random survival forest, implemented in R package DynForest, extends the joint modelling methodology to predict clinical events from individual longitudinal history when the number of repeated markers is large. Thanks to the random forests, it also naturally deals with complex relationships (interactions, nonlinear associations).

INVITED SESSION 6: Bayesian nonparametrics and ML for causal inference

IN6.1 Assumption-lean Cox regression

Stijn Vansteelandt[1], Torben Martinussen[2], Kelly Van Lancker[3], Oliver Dukes[4]

[1] Ghent University, Belgium, [2] University of Copenhagen, Denmark, [3] Johns Hopkins University, Baltimore MD, United States, [4] University of Pennsylvania, Philadelphia PA, United States

Inference for the parameters indexing Cox regression models is routinely based on the assumption that the model is correct and a priori specified. This is unsatisfactory because the chosen model is usually the result of a data-adaptive model selection process, which induces bias and excess uncertainty that is not usually acknowledged; moreover, the assumptions encoded in the resulting model rarely represent some a priori known, ground truth. Standard inferences may therefore lead to bias in effect estimates, and may moreover fail to give a pure reflection of the information that is contained in the data. Inspired by developments on assumption-free inference for so-called projection parameters, we here propose nonparametric definitions of main effect estimands which reduce to standard main effect parameters in Cox regression models when these models are correctly specified, but continue to capture the primary (conditional) association between a variable and an event time, even when these models are misspecified. We achieve an assumption-lean inference for these estimands by deriving their influence curve under the nonparametric model and invoking flexible data-adaptive algorithms. Insight will be given into the empirical performance of the proposal via simulation studies and a data analysis.

IN6.2 Adversarial Monte Carlo Meta-Learning of Conditional Average Treatment Effects

Alex Luedtke[1]

[1] *University of Washington, Seattle WA, United States*

We frame the meta-learning of conditional average treatment effect estimators as a search for an optimal strategy in a two-player game. In this game, Nature selects a prior over distributions that generate labeled data consisting of covariates, treatment, and an associated outcome, and the Estimator observes data sampled from a distribution drawn from this prior. The Estimator's objective is to learn a function that maps from a new feature to an estimate of the conditional average treatment effect. We establish that, under reasonable conditions, the Estimator's has an optimal strategy that is equivariant to shifts and rescalings of the outcome and is invariant to permutations of the observations and to shifts, rescalings, and permutations of the features. We introduce a neural network architecture that satisfies these properties.

IN6.3 Bayesian nonparametric methods for causal inference

Jason Roy[1]

[1] *Rutgers School of Public Health, Piscataway NJ, United States*

Bayesian nonparametric (BNP) methods can be used to flexibly model joint or conditional distributions, as well as functional relationships. These methods, along with causal assumptions for identification, can be used for inference about causal effects. This general approach to causal inference has several possible advantages over popular semiparametric methods, including efficiency gains, the ease of causal inference on any functionals of the distribution of potential outcomes, the use of prior information, and capturing uncertainty about causal assumption via informative prior distributions. In this overview talk we review approaches to causal inference problems using BNP methods, including modeling approaches based on Dirchlet process priors and Bayesian additive regression trees, for inference about average causal effects and mediation.

IN6.4 Nonparametric Bayesian Instrumental Variable Analysis: Evaluating Heterogeneous Effects of Coronary Arterial Access Site Strategies

Samrachana Adhikari[1], Sherri Rose[2], Sharon-Lise Normand[3]

[1] *NYU Grossman School of Medicine, New York NY, United States,* [2] *Stanford University, Stanford CA, United States,* [3] *Harvard Medical School and Harvard School of Public Health, Boston MA, United States*

Observational data introduces many practical challenges for causal inference. In this talk, I will focus on a particular issue when there are unobserved confounders such that the assumption of "ignorability" is violated. For making a valid causal inference in the presence of unmeasured confounders, instrumental variable (IV) analysis plays a crucial role. I will introduce a hierarchical Bayesian likelihood-based IV analysis under a Latent Index Modeling framework to jointly model outcomes and treatment status, along with necessary assumptions and sensitivity analysis to make a valid causal inference. The innovation in our methodology is an extension of existing parametric approach by i.) accounting for an unobserved heterogeneity via a latent factor structure, and ii.) allowing non-parametric error distributions with Dirichlet process mixture models. We demonstrate utility of our model in comparing effectiveness of two different types of vascular access for a cardio-vascular procedure.

INVITED SESSION 7: Semi-competing risks and causal inference

IN7.1 Estimands of practical interest in intercurrent event settings

Mats Stensrud[1]

[1] *Ecole Polytechnique Fédérale de Lausanne, Switzerland*

Investigators often express interest in treatment effects that suitably take account of post-treatment events (so-called intercurrent events). However, outcome contrasts that naively condition on intercurrent events do not have a straight-forward causal interpretation, and the practical relevance of other commonly used approaches is debated. 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 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 that were recently used to motivate consideration of principal stratum estimands in clinical trials. In all of these examples, I will suggest alternative causal estimands that correspond to explicit research questions of practical interest. These proposed estimands also require less stringent identification assumptions.

IN7.2 A Bayesian nonparametric approach for evaluating the causal effect of treatment in randomized trials with semi-competing risks

Daniel Scharfstein[1], Yanxun Xu[2], Peter Müller[3], Michael Daniels[4]

[1] *University of Utah School of Medicine, Salt Lake City UT, United States,* [2] *Johns Hopkins University, United States,* [3] *University of Texas, United States,* [4] *University of Florida, United States*

We develop a Bayesian nonparametric (BNP) approach to evaluate the causal effect of treatment in a randomized trial where a nonterminal event may be censored by a terminal event, but not vice versa (i.e., semi-competing risks). Based on the idea of principal stratification, we define a novel estimand for the causal effect of treatment on the nonterminal event. We introduce identification assumptions, indexed by a sensitivity parameter, and show how to draw inference using our BNP approach. We conduct simulation studies and illustrate our methodology using data from a brain cancer trial. The R code implementing our model and algorithm is available for download at https://github.com/YanxunXu/BaySemiCompeting.

IN7.3 A multistate approach for the study of interventions on an intermediate time-to-event in health disparities research

Linda Valeri[1], Cecile Proust-Lima[2], Weijia Fan[1], Jarvis T. Chen[3], Helene Jacqmin-Gadda[2]

[1] *Columbia University, New York NY, United States,* [2] *Universite de Bordeaux, France,* [3] *Harvard T.H. Chan School of Public Health, Boston MA, United States*

We propose a novel methodology to quantify the effect of stochastic interventions for a non-terminal intermediate time-to-event on a terminal time-to-event outcome. Investigating these effects is particularly important in health disparities research when we seek to quantify inequities in timely delivery of treatment and its impact on patients' survival time. Current approaches fail to account for time-to-event intermediates and semi-competing risks arising in this setting.
Under the potential outcome framework, we define causal contrasts relevant in health disparities research and provide identifiability conditions when stochastic interventions on an intermediate non-terminal time-to-event are of interest. Causal contrasts are estimated in continuous time within a multistate modeling framework and analytic formulae for the estimators of the causal contrasts are developed. We show via simulations that ignoring censoring in intermediate and/or terminal time-to-event processes, or ignoring semi-competing risks may give misleading results. This work demonstrates that rigorous definition of the causal effects and joint estimation of the terminal outcome and intermediate non-terminal time-to-event distributions are crucial for valid investigation of interventions and mechanisms in continuous time. We employ this novel methodology to investigate the role of delaying treatment uptake in explaining racial disparities in cancer survival in a cohort study of colon cancer patients.

PARALLEL SESSION 1: Adaptive designs

S1.1 Adaptive enrichment designs with a continuous biomarker

Nigel Stallard[1]

[1] *University of Warwick, Coventry, United Kingdom*

A popular design for randomised controlled clinical trials assessing targeted therapies is the two-stage adaptive enrichment design. In this design recruitment in the second stage is limited to a biomarker-defined subgroup chosen based on data from first stage. Data from both stages are then used to test a hypothesis concerning the treatment effect in the selected subgroup. The data-dependent selection leads to statistical challenges if data from both stages are used to draw inference on the treatment effect in the selected subgroup.
If the subgroups considered are nested, as when defined by a single continuous biomarker with an assumed non-decreasing treatment effect, treatment effect estimates in different subgroups follow the same distribution as estimates based on accumulating data in a group-sequential trial. This result is used to obtain a test that controls the familywise type I error rate in the wide range of settings with asymptotically normal test statistics. Two approaches are proposed; one is based directly on the multivariate normal distribution of the test statistics for the subgroups considered, and is suitable if the number of possible subgroups is small. The other is based on Brownian motion approximations and is suitable for selection of one of a large number of possible subgroups. The methods are illustrated using survival data from a breast cancer trial.

S1.2 A Bayesian multi-arm multi-stage design incorporating information about treatment ordering

Alessandra Serra[1], Pavel Mozgunov[1], Thomas Jaki[2]

[1] *MRC Biostatistics Unit, Cambridge, United Kingdom,* [2] *University of Cambridge, United Kingdom*

In this talk, we will present a Bayesian multi-arm multi-stage trial design that selects all the promising treatments with high probability and can efficiently incorporate the information about order in the treatment effects of the arms (e.g. when considering different treatment durations, different doses, or nested combination of treatments). A distinguishing feature of the design is that it allows taking into account the uncertainty of the treatment effect order assumption and it does not assume any parametric dose-response or duration-response model.
The focus of this talk will be on the implementation and evaluation of this design in a specific clinical trial setting. Specifically, we will cover how the decisions are made at each analysis and how the family-wise error rate and the power requirements are achieved. Via simulations, we will compare the proposed Bayesian design with the standard multi-arm multi-stage design and demonstrate the gains in the sample sizes the proposed design can provide. We demonstrate the robustness of the proposed design to violations of the assumptions on the order.

S1.3 Recommending a timing for a stop for efficacy in group sequential trials with a survival endpoint

Akane Yamakawa[1], Shoko Ukita[2], Shogo Nomura[3]

[1] *Foundation for Biomedical Research and Innovation at Kobe, Japan,* [2] *The University of Tokyo Graduate School of Interdisciplinary Information Studies, Japan,* [3] *Department of Biostatistics and Bioinformatics, Graduate School of Medicine, University of Tokyo, Japan*

Interim analyses are often planned and conducted in randomized clinical trials (RCTs) with long-term follow-up period. From our experiences and knowledge, the timing of interim analyses is routinely decided in two ways : based on the number of events (e.g., when 50%/75% of the planned required number of events is observed) or calendar time (e.g., when half/all of the required number of patients are enrolled). However, statistical properties of interim analysis timings are not fully discussed, especially for RCTs using a survival primary endpoint. The aim of this study is to discuss whether conventional timings of interim efficacy monitoring had favorable statistical properties. We defined “optimal timing” for interim analyses as the timing which minimizes ASN or average study period. All numerical calculation assumed that the calculation of the efficacy boundary was based on O’Brien-Fleming type alpha-spending function. For a setting with single efficacy monitoring and under proportionality of hazards, a change in ASN was negligible for almost all the scenarios. The optimal timing which minimized average study period ranged from 60% to 70% (information time scale). In a setting with two-times efficacy monitoring, the optimal timings were 50% to 60% and 70% to 80%. In the presentation, we also show several different performance when different statistical tests (e.g., G(ρ,γ) family tests or max-combo test) are used under non-proportionality of hazards situation.

S1.4 Two-stage designs for clinical trials with small sample sizes

Meinhard Kieser[1], Nico Bruder[1], Philip Winkler[1], Meis Jan[1], Maximilian Pilz[1]

[1] *Institute of Medical Biometry, University of Heidelberg, Germany*

When applying group-sequential designs in clinical trials with normally distributed outcomes, approximate critical values are often applied. Here, normally distributed test statistics are assumed which, however, are in fact t-distributed. For small sample sizes, the approximation may lead to a serious inflation of the type I error rate. Recently, Rom and McTague (2020) proposed a method for computing the exact critical boundaries assuring type I error rate control. They provided the critical boundaries for Pocock- and O’Brien-Fleming-like group-sequential designs. For designs with one interim analysis, we present six alternative designs, which also control the type I error rate and in addition allow flexible design modifications. We compare the characteristics of these seven two-stage designs. It is shown that considerable sample size savings can be achieved by including futility stopping and by optimizing the designs. A freely available software package is presented where these methods are implemented In summary, for clinical trials with small sample sizes as, for example, in the area of rare diseases, optimal two-stage designs with futility stopping may be a valuable alternative to classical group-sequential designs.

REFERENCE:

Rom, D. M., and J. A McTague. 2020. Exact critical values for group sequential designs with small sample sizes. Journal of Biopharmaceutical Statistics 30:752–764. doi: 10.1080/10543406.2020.1730878.

Parallel Sessions

S1.5 Cross-validated risk scores adaptive enrichment design

Svetlana Cherlin[1], James Wason[1]

[1] *Newcastle University, Newcastle upon Tyne, United Kingdom*

BACKGROUND: Adaptive enrichment clinical trial designs allow the trial to update the inclusion criteria based on the interim analysis. In the second stage, the entry is restricted to the subgroup of patients who are predicted to benefit from the treatment. Current adaptive enrichment methods assume that the subgroup is defined by a known predictive biomarker, which might not be available. With the recent advances in multi-omics technologies, increasingly large numbers of biomarkers are becoming available. Several approaches that utilise high-dimensional data have been proposed, such as the risk scores approach that summarises the high-dimensional information into a single score for each patient. The risk scores are subsequently used for identifying a subgroup of patients who benefit from the treatment.
METHOD: We propose a design that allows enriching the recruitment with patients who are predicted to benefit from the treatment, based on their high-dimensional baseline covariates. The sensitive group is identified using the risk score approach where each patient is assigned a score constructed from their baseline covariates. The design includes early stopping for futility if no promising treatment effect is identified in the sensitive group and also the difference between the arms in the overall trial population is not significant. We have implemented the new methods in an R package.
RESULTS: We have investigated the performance of the proposed design by applying it to simulated data scenarios with various response rates for the sensitive group and different sample sizes. The design allows to narrow down the eligibility and also achieves this at a smaller expected sample size, in comparison to the non-enrichment alternative. For the null scenario, the design achieves a well-controlled type I error rate with a substantial reduction in the expected sample size (at least 24%). We illustrated our approach on a randomised clinical trial with publicly available high-dimensional gene expression data.
CONCLUSIONS: The new method shows a superior performance in terms of the power and the sample size in comparison to the non-enrichment approach. Further work could explore different distributions of outcomes, as well as multiple arms and endpoints.

PARALLEL SESSION 2: Causal Inference

S2.1 Using many invalid instrumental variables to tighten inference on causal effects

Ashish Patel[1], Dipender Gill[2], Paul Newcombe[1], Stephen Burgess[3]

[1] *MRC Biostatistics Unit, Cambridge, United Kingdom,* [2] *Imperial College London, United Kingdom,* [3] *University of Cambride, United Kingdom*

We consider the problem of improving inference in instrumental variable models when only a few instruments are plausibly valid, while many others may be invalid. This problem is motivated by the widely-used method of Mendelian randomization (MR), which is used to identify causal links between a risk factor and disease. MR involves choosing appropriate genetic variants as instruments, in a bid to remove bias from unobserved confounding. Although many genetic variants may be invalid instruments (due to their direct effects on the outcome which are not through the exposure), their inclusion could increase the precision of causal effect estimates at the cost of allowing some bias. We explore how to optimally tackle this bias-variance trade-off by carefully choosing from many weak and locally invalid instruments.
Specifically, we consider a "focused" selection approach for a two-sample summary data setting, where instruments are selected on the basis of how they impact the asymptotic mean square error of causal effect estimates. We show how less restrictive assumptions on direct effects allow for only conservative model selection. We then propose a novel method to tighten honest confidence intervals through support restrictions on invalid instruments, which may be motivated through weak assumptions used in the partial identification literature. When applying our results to MR studies instrumenting for vitamin D and systolic blood pressure, our findings suggest that optimal model selection does not involve only a small number of biologically-justified instruments, but additionally hundreds of potentially invalid instruments.

S2.2 Sensitivity Analysis of Causal Effects in Observational Studies

Md. Abdul Basit[1], AHM Mahbub Latif[1]

[1] *Institute of Statistical Research & Training, University of Dhaka, Bangladesh*

In observational studies, the identification of causal estimands of interest depends on the identifiability assumptions, and one of the most common identifiability assumptions is the no unmeasured confounding (NUC) assumption. The NUC assumption is not verifiable from observed data, and violation of this assumption may induce systematic bias in the causal effect estimates. Sensitivity analysis is conducted to assess the robustness of estimated causal effects to the violation of the NUC assumption in observational studies. This paper proposes two extensions of a recent sensitivity analysis framework proposed by Zhao et al. (2019) for binary treatments, which quantifies the magnitude of unmeasured confounding using odds ratios between the unobserved and observed propensity scores. Since risk ratios are easier to interpret and are consistent with general intuition, a risk ratio (RR) based sensitivity analysis framework is proposed. Furthermore, recent years have seen a growing interest in developing causal inference methods for observational studies with multivalued treatments. However, almost all of the sensitivity analysis frameworks found in causal inference literature to date are applicable to binary treatments only. Observing this scarcity of sensitivity analysis frameworks for multivalued treatments, the proposed RR framework is extended to the multivalued treatment setting using generalized propensity scores. Simulation studies are conducted to evaluate the performance of the proposed sensitivity analysis frameworks in terms of the bias in the point estimate intervals and the non-coverage of the percentile bootstrap confidence intervals. Simulation results suggest that the proposed RR-based sensitivity analysis frameworks perform well in both binary and multivalued treatment settings when there is an adequate overlap in the covariate distribution among the treatment groups. Lastly, an R package is developed that can be used to conduct sensitivity analysis in observational studies with both binary and multivalued treatments using the proposed frameworks.

S2.3 Investigating inequalities in cancer survival through mediation analysis with limited interventions

Sarah Booth[1], Elisavet Syriopoulou[2], Mark Rutherford[1], Paul Lambert[1,2]

[1] *University of Leicester, United Kingdom*
[2] *Karolinska Institutet, Solna, United Kingdom*

Cancer survival varies substantially by socioeconomic status. This is partly due to the timing of diagnosis, as patients in the most deprived groups tend to be diagnosed at a later stage with more advanced cancer.
Mediation analysis, using deprivation as the exposure and stage of tumour at diagnosis as the mediator, can be used to understand to what extent differences in survival are due to factors relating to socioeconomic status and what proportion can be explained by differences in the stage distribution. This partitions the total difference in survival across deprivation groups into the direct effect of deprivation on cancer survival and the indirect effect of deprivation via the mediator (stage differences).
Whilst survival differences relating to socioeconomic factors may not be able to be removed completely, we can hypothesise interventions that may eliminate the indirect effect of deprivation by achieving earlier diagnosis amongst the most deprived group of patients. In this work, using regression standardization, we examine the improvements in survival and gains in life expectancy if the most deprived group could have been diagnosed earlier such that they had the same stage distribution as the least deprived group.
However in practice, the differences in the stage distributions may only be able to be partially eliminated as it may not be feasible to completely achieve the stage distribution of the least deprived group with a given intervention. Therefore in this work, we also investigate the effect that more realistic, smaller shifts in the stage distribution have on survival and the number of life years gained for the most deprived group. We examine different ways of quantifying the shift in stage distribution and advocate performing a sensitivity analysis over a range of realistic scenarios.

S2.4 Inferring the Effect Direction in Genetic Association Studies

Sharon Lutz[1], Stijn Vansteelandt[2], Christoph Lange[3]

[1] Harvard Medical School, Boston MA, United States, [2] Ghent University, Belgium, [3] Harvard T.H. Chan School of Public Health, Boston MA, United States

In genetic association studies, Mendelian Randomization (MR) has gained in popularity as a concept to assess the causal relationship between two phenotypes. Recent methods have been proposed as tools that can infer the causal direction between two phenotypes: MR Steiger, bidirectional MR, causal direction-ratio, causal direction-Egger, and causal direction-GLS. Through simulation studies, we examined the ability of these 5 approaches to correctly determine the effect direction in the presence of pleiotropy, measurement error, unmeasured confounding, and weak instrument variables. In addition, we examined the performance of these approaches when there is a longitudinal causal relationship between the two phenotypes, differing distributions for the phenotypes, and selection bias in the study design. We also applied these methods to the UK biobank and COPDGene study to examine the role of smoking on pulmonary function in the presence of possible pleiotropy, measurement error, and/or selection bias.

S2.5 Personalized Biopsy Schedules Using Cause-specific Interval-censored Joint Models

Zhenwei Yang[1], Dimitris Rizopoulos[1], Eveline Heijnsdijk[1], Lisa Newcomb[2], Nicole Erler[1]

[1] Erasmus Medical Center Rotterdam, Netherlands
[2] Fred Hutchinson Cancer Research Center, Seattle WA, United States

BACKGROUND: Prostate cancer is often slow-growing and may not need immediate treatment. Patients in early stages are admitted to active surveillance programs where biopsies are conducted regularly to detect cancer progression. The routine use of such invasive tests, however, burdens the patients. An emerging proposal is to replace the fixed biopsy schedules with flexible schedules based on each patient's risk of progression. This risk is estimated using repeatedly measured Prostate-Specific Antigen (PSA) levels and the results and timing of previous biopsies. Periodical biopsies prevent the time of progression from being observed exactly, leading to interval censoring. Furthermore, the natural history to progression can be interrupted by initiating treatment, consitituting a competing risk and complicating the setting further. In the Canary Prostate Active Surveillance Study (PASS) that motivates our work, this is the case for ~10% of the patients. To obtain optimal personalized biopsy schedules, these data characteristics have to be modeled appropriately, requiring a new methodology.
OBJECTIVE:
1) To extend the joint modeling framework to competing risk settings with both interval and right censored events.
2) Development of personalized biopsy schedules that reduce the number of biopsies while limiting the delay in diagnosis of progression.
METHOD: We develop a Bayesian cause-specific interval-censored joint model that allows us to derive the progression-specific risk for new patients based on their trajectories of PSA levels. The underlying likelihood function incorporates the interval-censoring of cancer progression, the competing risk of treatment, and the uncertainty about whether cancer progression occurred since the last negative biopsy in patients that are right-censored or experience the competing event. Future biopsies are scheduled when the estimated progression-specific risk exceeds a certain threshold. We also offer a methodology to choose the optimal threshold based on a loss function balancing the expected number of biopsies and the delay of detection.
RESULTS: Preliminary results from the analysis of the 833 patients in the PASS data show that our personalized schedules reduce the expected number of biopsies by 31% per patient with expected detection delay decreasing by 0.35 years compared to a fixed biennial schedule.

PARALLEL SESSION 3: Missing data

S3.1 Multiple imputation for the Fine-Gray model: an approach based on the subdistribution process

Edouard F Bonneville[1], Jan Beyersmann[2], Johannes Schetelig[3], Hein Putter[1], Hoogleraar Liesbeth C de Wreede[1,4]

[1] Department of Biomedical Data Sciences, Leiden University Medical Center, Netherlands, [2] Institute of Statistics, Ulm University, Germany, [3] Dresden University Hospital, Germany, [4] DKMS, Dresden/Tübingen, Germany

The Fine-Gray (FG) model is increasingly used to build prognostic models in the presence of competing events. When missing values occur in covariates during model development, researchers often choose to multiply impute them. Specific guidance on multiple imputation (MI) in the FG context, such as what predictors need to be included in the imputation model, is scarce. Indeed, such an imputation model can be difficult to motivate: it requires assumptions on the competing failure causes (due to the form of the full data likelihood), which the FG model by design does not make or help inform. In this work, we motivate an imputation model in the FG context from the point of view of the subdistribution process. Specifically, we base the fully conditional distribution of the missing covariate(s) given outcome and observed covariates on the likelihood for the subdistribution time for the event of interest. We show that in the presence of right-censoring, such an imputation model can be implemented by also imputing the censoring times for individuals failing from competing events. This leads to so-called 'censoring complete data', allowing the use of standard software for fitting the analysis model in the imputed datasets. The performance of the proposed method is evaluated in a simulation study, where misspecification of the underlying data-generating model is of particular interest. That is, a) assessing in a first instance how the proposed method performs when the FG model holds for the cause of interest; and b) assessing the performance when simulated data follow proportional cause-specific hazards. Furthermore, the proposed approach is applied on a dataset from the field of hematopoietic stem cell transplantation. Results are discussed in the context of earlier work on MI for cause-specific Cox models.

S3.2 Record linkage with complex correlation structures

Michel Hof[1]

[1] Amsterdam University Medical Center, Netherlands

With the rise of digitization, an increasing amount of data is electronically stored and it is becoming more and more common to combine existing data to answer new research questions. In absence of a unique identifier (e.g. citizen service number), combining the data from multiple resources often relies on partially identifying variables (e.g. gender, place of residence, and initials). With a record linkage procedure, these variables are used to distinguish pairs of observations that belong together (matches) from pairs of observations that do not belong together (non-matches).
A natural approach to record linkage is to treat the binary matching indicators as missing data and fit a marginal model on the observed values of the partially identifying variables. This approach, however, is computationally unfeasible as it involves a very high dimensional integral over the matching indicators from all observations pairs. As a solution, many record linkage approaches have been proposed that (incorrectly) assume independence between the matching indicators. Complex correlation structures (e.g. each observation can maximally form one match with an observation in the other dataset) are ignored, which increases the risk of incorrect identification of matches and non-matches. Recently, a number of Bayesian approaches have been proposed to deal with these complex correlation structures. Although the results of these approaches are promising in terms of identifying matches, they are only computationally feasible with relatively small datasets (e.g. containing not more than 5000 observations). In practice, it is common to encounter datasets with far more observations.
In this work, we propose an alternative approximation method based on an expectation maximization (EM) algorithm. To lower the computational burden, we avoid calculating the marginal likelihood by using Monte Carlo approximation methods. Simulations showed that the model has similar performance as the Bayesian approaches in terms of identifying matches and non-matches. To illustrate that the new model is scalable to huge datasets, we fitted the new model to the Perinatal Registry of the Netherlands to identify first (n=600000) and second born (n=400000) children that belong to the same mother.

S3.3 Quantitative bias analysis for unmeasured confounding: Review of software for regression

Emily Kawabata[1], Kate Tilling[1], Rolf Groenwold[2], Rachael Hughes[1]

[1] University of Bristol, United Kingdom, [2] Leiden University Medical Center, Netherlands

Failure to appropriately account for unmeasured confounding in analyses may lead to bias and erroneous conclusions. Quantitative bias analysis (QBA) for unmeasured confounding is used to quantify the potential direction and impact of the bias or to quantify how much unmeasured confounding would be needed to change a study's conclusions. Generally, a QBA will include one or more parameters (known as bias parameters) which cannot be identified from the observed data. Information on plausible values for these bias parameters can be obtained from external sources (such as published literature) or from calibration of the bias parameters to the measured confounders. The adoption of QBA by applied researchers has been slow, partly due to the focus of methods for binary outcomes and exposures, and partly due to the lack of accessible software. We provide a review of the latest developments in QBA software between 2011 to 2021. We then provide a detailed comparison of five different QBA methods, and their software implementations, applicable when the analysis of interest is a linear regression. The software implementations we compare are treatSens, causalsens, sensemakr, E-value, and konfound and we illustrate application of these software implementations to two real datasets. A panel of applied researchers provide their feedback on the accessibility of these five methods. Our review reported many new QBA software implementations since the last published review, most of which are deterministic QBA methods implemented in the freely available statistical software environment R. Many programs include features such as calibration and graphical displays of the QBA results to aid interpretation. Our comparative evaluation illustrated the wide variation in the types of QBA methods applicable to a linear regression analysis. QBA methods where the bias parameters are easy to interpret and routinely reported in the published literature are preferable. However, QBA methods that solely assess sensitivity to a change in statistical significance should be discouraged. In summary, the diversity of QBA methods presents challenges to the widespread uptake of QBA among applied researchers. Provision of detailed guidelines and software implementations in platforms other than R would be beneficial.

S3.4 Sensitivity analysis for calibrated weighted estimators under non-ignorable dropout

Li Su[1], Shaun Seaman[1], Sean Yiu[1]

[1] University of Cambridge, United Kingdom

Inverse probability of censoring weighting is a popular approach to handling dropout in longitudinal studies. However, inverse probability-of-censoring weighted estimators (IPCWEs) can be inefficient and unstable if the weights are estimated by maximum likelihood. To alleviate these problems, calibrated IPCWEs have been proposed, which use calibrated weights that directly optimize covariate balance in finite samples rather than the weights from maximum likelihood. However, the existing calibrated IPCWEs are all based on the unverifiable assumption of sequential ignorability and sensitivity analysis strategies under non-ignorable dropout are lacking. In this paper, we fill this gap by developing an approach to sensitivity analysis for calibrated IPCWEs under non-ignorable dropout. A simple technique is proposed to speed up the computation of bootstrap and jackknife confidence intervals and thus facilitate sensitivity analyses. We evaluate the finite-sample performance of the proposed methods using simulations and apply our methods to data from an international inception cohort study of systemic lupus erythematosus. An R Markdown tutorial to demonstrate the implementation of the proposed methods is provided.

S3.5 Standard and reference-based imputation methods based on conditional mean imputation

Marcel Wolbers[1], Alessandro Noci[1], Paul Delmar[1], Craig Gower-Page[2], Sean Yiu[2], Jonathan Bartlett[3]

[1] Data & Statistical Sciences, Pharma Development, Roche, Basel, Switzerland, [2] Data & Statistical Sciences, Pharma Development, Roche, Welwyn Garden City, United Kingdom, [3] Department of Mathematical Sciences, University of Bath, United Kingdom

Many randomized controlled clinical trials compare a continuous outcome that is assessed longitudinally at scheduled follow-up visits between subjects assigned to an intervention treatment group and those assigned to a control group. Missing outcome measurements may occur because subjects miss an assessment or drop out from the trial altogether. Moreover, intercurrent events (ICEs) such as discontinuations of the assigned treatment or initiations of rescue medications may affect the interpretation or the existence of the outcome measurements associated with the clinical question of interest. The ICH E9(R1) addendum on estimands presents a structured framework to link trial objectives to a precise description of the targeted treatment effect in the presence of ICEs and missing data.
Missing data methods based on multiple imputation are increasingly used to align the analysis strategy with the defined estimand. Imputations may be based on a missing-at-random assumption or on a reference-based imputation assumption. Reference-based methods impute missing data in the intervention treatment group based on observed data from a reference group which is typically defined as the control group of the trial. Conventionally, imputation is implemented using Bayesian random multiple imputation and Rubin's rules for pooling results across imputed datasets. However, this approach requires the specification of prior distributions and MCMC sampling. Moreover, it overestimates the frequentist standard error for reference-based imputation.
We propose and justify deterministic conditional mean imputation (based on maximum likelihood estimation of imputation parameters) combined with the jackknife for inference as an alternative approach. In an application and a simulation study, we demonstrate that our proposal provides unbiased treatment effect estimates and correct frequentist inference with accurate standard error estimation and type I error control. Additionally, it can result in substantially more efficient treatment effect estimators under reference-based imputation assumptions than the Bayesian approaches. A further advantage of the method is that it does not rely on random sampling and is therefore easily replicable and unaffected by Monte Carlo error. The implementation of the method in the publicly available R package "rbmi" will also be described.

PARALLEL SESSION 4: Survival data

S4.1 General independent censoring in event-driven trials with staggered entry

Jasmin Rühl[1], Jan Beyersmann[2], Sarah Friedrich[1]

[1] University of Augsburg, Germany, [2] Institute of Statistics, Ulm University, Germany

Randomised clinical trials with time-to-event endpoints are frequently stopped after a pre-specified number of events has been observed. This practice leads to dependent data and non-random censoring, though, which can generally not be solved by conditioning on the underlying baseline information. If the observation period starts at the same time for all subjects, the assumption of independent censoring in the counting process sense is valid. However, in the (more common) case of staggered study entry, matters are complicated substantially, and a formal proof that justifies the use of the common time-to-event methods has not yet been given.
We demonstrate that the study design at hand still entails general independent censoring, provided that the analysis is based on study time information only. To illustrate that the filtrations must not use abundant information, we simulated event-driven trials with staggered entry and evaluated them by means of Cox regression models with covariates for the calendar times. The associated Breslow curves of the cumulative baseline hazard showed considerable deviations. This implies that the analysis is disturbed by the condition on calendar time variables. In addition, we present the results of a simulation study that highlights the need for caution when using methods based on random censoring: Efron's classical bootstrap provided biased results in the given setting, whereas the accuracy of the (martingale-based) wild bootstrap was not affected.

Parallel Sessions

S4.2 Follow-up time in clinical trials with a time-to-event endpoint: Redefining the question(s)

Kaspar Rufibach[1], Lynda Grinsted[2], Jiang Li[3], Yue Shentu[4], Hans-Jochen Weber[5], Cheng Zheng[6], Jiangxiu Zhou[7]

[1] F. Hoffmann-La Roche, Basel, Switzerland, [2] AstraZeneca, [3] BeiGene, [4] Daiichi Sankyo, [5] Novartis, [6] Zentalis, [7] GSK

For the analysis of a time-to-event (T2E) endpoint in a single-arm (SAT) or randomized clinical trial (RCT) it is generally perceived that interpretation of a given estimate of the survival function, or the comparison between two groups, hinges on some quantification of the amount of follow-up. Typically, a median of some loosely defined quantity is reported. However, as already discussed in Shuster in 1991, whatever median is reported is typically not answering the question(s) trialists actually have in terms of follow-up quantification. This talk will focus on the following discussion points:
- The estimand framework put forward in the ICH E9 estimand addendum has been broadly implemented in pharmaceutical drug development, so far primarily for efficacy endpoints. We have found that taking inspiration from the estimand framework to structure the question of follow-up for a T2E endpoint and all the quantities that have been proposed to «estimate» it allows for a transparent way of describing and appreciating the various approaches.
- Following the thinking process in the addendum we formulate a comprehensive list of relevant scientific questions that trialists have when reporting T2E data and which are often «answered» with reference to some unclearly defined quantifier of follow-up. We illustrate how instead these questions should be answered, and that reference to an unclearly defined «follow-up quantity» is not necessary.
- The literature so far has focused on quantifying follow-up for estimation of a survival function in one group. However, in oncology drug development key decisions are made based on RCTs, and we discuss relevant scientific questions in this context.
- Although our conclusion will be that generally used follow-up quantifiers are not useful, we define and illustrate follow-up quantifiers, some of which have not been discussed in the statistical literature so far but are routinely used in reporting of trials. Finally, with the advent of immunotherapies in oncology patterns of survival functions in RCT emerged, e.g. delayed separation, that may require different thinking on some of the relevant scientific questions.

S4.3 Stratified modestly-weighted log-rank tests in settings with delayed separation of survival curves

Dominic Magirr[1], Jose L. Jimenez[1]

[1] Novartis, Riehen, Switzerland

Delayed separation of survival curves is a common occurrence in confirmatory studies in immuno-oncology. Many novel statistical methods that aim to efficiently capture potential long-term survival improvements have been proposed in recent years. However, the vast majority do not consider stratification, which is a major limitation considering that most large confirmatory studies currently employ a stratified primary analysis. In this article, we combine recently proposed weighted log-rank tests that have been designed to work well under a delayed separation of survival curves, with stratification by a baseline variable. The aim is to increase the efficiency of the test when the stratifying variable is highly prognostic for survival. As there are many potential ways to combine the two techniques, we compare several possibilities in an extensive simulation study. We also apply the techniques retrospectively to two recent randomized clinical trials.

S4.4 Restricted mean survival time regression model with time-dependent covariates

Chengfeng Zhang[1], Baoyi Huang[1], Hongji Wu[1], Yawen Hou[2], Zheng Chen[1]

[1] Southern Medical University, Guangzhou, China, [2] Jinan University, Guangzhou, China

In clinical or epidemiological follow-up studies, methods based on time scale indicators such as the restricted mean survival time (RMST) have been developed to some extent. Compared with traditional hazard rate indicator system methods, the RMST is easier to interpret and does not require the proportional hazard assumption. To date, regression models based on the RMST are indirect or direct models of the RMST with baseline covariates. However, time-dependent covariates are becoming increasingly common in follow-up studies. To address survival data with time-dependent covariates and time-effect covariates, we first proposed time-dependent RMST (T-RMST) regression based on the IPCW method, which is handled exogenous time-dependent covariates. Further to deal with endogenous time-dependent covariates (longitudinal covariates), we proposed joint models of RMST and longitudinal covariates (JM-RMST) on the basis of the T-RMST models, which can dynamically predict individual survival time. Through Monte-Carlo simulation, We verified the estimation performance of the regression parameters of the T-RMST model, and the prediction performance is better than time-dependent Cox and fixed covariate RMST regression. Besides, JM-RMST regression coefficient estimates are also stable, and the prediction performance is better than the fixed covariate RMST regression. Finally, time-dependent RMST regression and JM-RMST regression are illustrated by two examples.

S4.5 Efficient estimation of joint models for multivariate longitudinal and survival data using INLA

Denis Rustand[1], Janet van Niekerk[1], Elias Teixeira Krainski[1], Haavard Rue[1], Cécile Proust-Lima[2]

[1] King Abdullah University of Science and Technology, Thuwal, Saudi Arabia, [2] Inserm BPH U1219, Université de Bordeaux, France

Joint models for longitudinal and survival outcomes have recently gained a lot of interest in clinical research. These complex models involve multiple likelihoods (i.e. for each longitudinal and survival outcome), usually linked through correlated or shared random-effects. In this context, inference methods reach limitations due to long computation times and convergence issues. We introduce a Bayesian approximation for these joint models based on the INLA algorithm implemented in the R package INLA to alleviate the computational burden and allow the estimation of multivariate joint models with less restrictions. Our simulation studies show that INLA reduces the computation time substantially as well as the variability of the parameter estimates compared to alternative strategies such as Bayesian inference via Markov Chain Monte Carlo or Maximum likelihood estimation with Monte Carlo expectation maximisation. We further apply the methodology to analyze 5 longitudinal markers (3 continuous, 1 count, 1 binary, and 16 random effects) and competing risks of death and transplantation in a clinical trial on primary biliary cholangitis. INLA provides a fast and reliable inference technique for applying joint models to the complex multivariate data encountered in health research.

PARALLEL SESSION 5: Software Engineering

S5.1 Introduction from the Academic Perspective

Martin Shaw[1]

[1] *University of Glasgow, United Kingdom*

S5.2 Introduction from the Industry Perspective

Daniel Sabanes Bove[1], Heidi Seibold[2]

[1] *Hoffmann-La Roche Ltd, Basel, Switzerland,* [2] *Johner Institut GmbH*

S5.3 PANEL DISCUSSION

Armin Schueler[1], Andy Nicholls[2], Anne-Laure Boulesteix[3], Alessandro Gasparini[4], Martin Shaw[5], Daniel Sabanes Bove[6]

[1] *Merck Healthcare,* [2] *GSK,* [3] *LMU Munich, Germany,* [4] *Karolinska Institutet, Sweden,* [5] *University of Glasgow, United Kingdom,* [6] *Hoffmann-La Roche Ltd, Basel, Switzerland*

The term "Research Software Engineer" (RSE) was coined by a group of UK software developers working in research in 2012. Since then grass-roots movements have sprung up all over the world with the aim of improving research software engineering and establishing career paths for RSEs. Following the motto "Better Software, Better Research", we aim to use this session to discuss the role of research software engineering in clinical biostatistics, both in academia and in industry.
After short input talks from the two perspectives of academia and industry, we invite for an interactive discussion with a small panel and the session participants to discuss (among others) the following questions:
(1) How to professionalize software engineering in clinical biostatistics?
(2) How can RSEs help with reproducibility of statistical analyses?
(3) How to start new software projects across academia and industry?
(4) How to create long term career paths for RSEs in clinical biostatistics?

PARALLEL SESSION 6: Cluster Trials

S6.1 Decaying correlation parameter values obtained from previously analysed cluster randomised trials

Jessica Kasza[1], Rhys Bowden[1], Yongdong Ouyang[2], Monica Taljaard[2], Andrew Forbes[1]

[1] *Monash University, Melbourne, Australia,* [2] *Ottawa Hospital Research Institute, Ottawa ON, Canada*

A frequently-applied assumption in the analysis of data from cluster randomised trials is that the outcomes from all participants within a cluster are equally correlated. That is, the intracluster correlation, which describes the degree of dependence between outcomes from participants in the same cluster, is the same for each pair of participants in a cluster. However, recent work has discussed the importance of allowing for this correlation to decay as the time between the measurement of participants in a cluster increases. Incorrect omission of such a decay has implications for the design and analysis of cluster randomised trials: for example, confidence intervals for estimated treatment effects can be too narrow or too wide, depending on the characteristics of the design. When planning studies, researchers often rely on previously reported analyses of trials to inform their choice of intracluster correlation. While some estimates of decaying correlations are available, most reported analyses of clustered data do not incorporate a decay. Researchers planning a trial under a decaying correlation may then face the challenge of specifying plausible combinations of decaying correlation parameters that are compatible with the intracluster correlation value obtained under no decay.
In this talk we show that it is possible to use intracluster correlation values obtained from models that incorrectly omit a decay to inform plausible choices of decaying correlations. Our focus is on intracluster correlation estimates for continuous outcomes obtained by fitting linear mixed models with exchangeable or block-exchangeable correlation structures. We describe how plausible values for decaying correlations (i.e. those correlations associated with discrete time decay correlation structures) may be obtained given these estimated intracluster correlations. We will demonstrate an online app that allows users to obtain these consistent values. This can be used at the trial planning stage to assess the sensitivity of sample size and power calculations to decaying correlation structures.

S6.2 The staircase cluster randomised trial design: a pragmatic alternative to the stepped wedge design

Kelsey Grantham[1], Andrew Forbes[1], Jessica Kasza[1]

[1] *Monash University, Melbourne, Australia*

The stepped wedge cluster randomised trial design, which randomises the order in which clusters of participants implement the intervention, is becoming increasingly popular. A staircase design is a variant of the stepped wedge design, named for the zig-zag pattern of steps along the diagonal of a stepped wedge design schematic where clusters switch from control to intervention conditions. Unlike a complete stepped wedge design where all participating clusters must collect and provide data for the entire trial duration, clusters in a staircase design are only required to be involved and collect data for a limited number of pre- and post-switch periods. This could alleviate some of the burden on participating clusters, encouraging involvement in the trial and reducing the likelihood of attrition.
Why the staircase? It has been shown that not all measurements in a complete stepped wedge design contribute equally towards estimation of the treatment effect. Specifically, measurements from clusters in periods just before or after the switch from the control to the intervention conditions contribute more information for estimation of the treatment effect than many of the other cluster-period cells. These “information-rich” cells form the basis of the staircase design. In this talk we describe the properties and potential advantages of the staircase design over the complete stepped wedge design. We present scenarios under which the staircase design is a more practical and/or more efficient alternative to the complete stepped wedge design. We also explore the relative efficiency of the designs under different assumed within-cluster correlation structures and for different trial configuration parameters likely to be seen in practice.

Parallel Sessions

S6.3 What is the point of point estimates?

Stephen Senn[1]

[1] *Statistical Consultant Edinburgh, United Kingdom*

It is now nearly 100 years since RA Fisher proposed randomisation as a means to enable precise statements to be made about inevitably imprecise inferences. Probability was always part of his program. However, point estimates on their own have no element of uncertainty attached to them and the common obsession with point estimates and the accompanying concept of bias regularly confuses, not to say confounds, discussions of randomisation.
Common misunderstandings of clinical trials include that randomisation is meant to guarantee balance, that large trials are more balanced than small ones and that there are potentially infinitely many confounders and that therefore randomisation is pointless. It is depressing to find how often these claims are made and that even statisticians make them.
The currently fashionable (but also powerful and important) field of causal inference also stresses point estimates and identifiability rather then (say) confidence or credible intervals and estimability and this is hampering correct handling of complex data-sets.
I shall argue that we (statisticians) need to get back to basics on understanding of randomisation and components of variance in order to communicate these better. Ironically a powerful tool for doing this is the Rothamsted approach to analysing designed experiments, one which has long been available to agricultural scientests but which has been largely ignored by medical statisticians.
Poperly understood, point estimates are degenerate confidence intervals in which we have zero confidence. Grasping this simple point would help understanding of what randomisation can and cannot achieve but also as to why cuasal inference needs to pay attention to standard errors.

S6.4 Ratio-of-ratio estimator of direct intervention effect in cluster-randomized trials

Yin Bun Cheung[1], Xiangmei Ma[1], KF Lam[2], Paul Milligan[3]

[1] *Duke-NUS Medical School, Singapore,* [2] *University of Hong Kong,* [3] *London School of Hygiene & Tropical Medicine, United Kingdom*

We propose and evaluate a ratio-of-ratio estimator of direct intervention effect on event rates in cluster-randomized trials where only a specific group of cluster members are eligible for the interventions. For example, in trials of vaccines against paediatric infectious diseases, it is common that only infants/young children in a specific age range are eligible to receive the vaccines or their comparators. Older children/adults are not. They are referred to as the target and non-target groups here, respectively. Within each trial arm, the ratio of event counts in the target group to that in the non-target group is estimated. Then, the ratio of the event count ratio in the intervention trial arm to that in the control trial arm is obtained. We show that this estimator is asymptotically unbiased for the estimation of the direct intervention effect. We also provide a bias-corrected version of this estimator for use in situations of small number of clusters. Importantly, we show that the ratio-of-ratio estimators are more powerful than popular alternatives when the clusters are highly heterogeneous (i.e. large coefficient of variation in event rates), because the use of data from the non-target group is noise-cancelling. We discuss the role of the estimators in study planning and analysis.

S6.5 Estimating the intervention effect using restricted mean survival time in a cluster randomized trial

Floriane Le Vilain-Abraham[1], Elsa Tavernier[1], Etienne Dantan[2], Solène Desmée[1], Agnès Caille[1]

[1] *Université de Tours, France; INSERM, SPHERE U1246,* [2] *Université de Nantes, France; INSERM, SPHERE U1246*

INTRODUCTION: In randomised clinical trials with a time-to-event outcome, the difference in restricted mean survival time (DRMST) between the intervention and control arms is an alternative summary measure of the intervention effect, not relying on the proportional hazards assumption. The DRMST is easily interpretable as the expected survival duration gained due to intervention over t*. To our knowledge, the estimation of a DRMST has not been explored in the specific context of cluster randomized trials (CRTs), in which social units are randomized to intervention or control conditions, thus introducing a correlation between the survival times of the subjects of a same cluster. We aimed to compare several methods to estimate a DRMST in CRTs and assess their statistical performances by using a simulation study.
METHODS: We extended existing methods for independent survival data to the CRT framework. The first method relies on the direct integration of Kaplan-Meier curves (method 1). We accounted for the clustering by using a bootstrap variance estimate (method 2). The second approach is the pseudo-values regression, which consists of computing pseudo-values for each individual and considering them as the response in a generalized linear model fitted with generalized estimating equations (method 3). We adjusted for clustering with a robust sandwich variance and either an exchangeable (method 4) or an independent working correlation matrix (method 5). We compared the 5 methods (1 and 3 not adjusting for clustering and 2, 4 and 5 adjusting for clustering) by using a simulation study with several scenarios (number of clusters, cluster size, intervention effect, degree of clustering, proportional and non-proportional hazards).
RESULTS: The variance of the DRMST was underestimated for methods 1 and 3, thus leading to uncontrolled type I error rate. Methods 2, 4 and 5 well estimated the variance and well controlled the type I error rate when the total number of clusters was large enough (≥50). The three methods adjusting for clustering performed similarly. The results were qualitatively similar under the proportional and non-proportional hazards patterns. We illustrated the proposed methods by using a CRT evaluating the effectiveness of an asthma control education program.
CONCLUSION: This work opens the way for better handling time-to-event outcomes in CRTs.

PARALLEL SESSION 7: High dimensional data

S7.1 Rank-based Bayesian variable selection for genome-wide transcriptomic analyses

Emilie Eliseussen Ødegaard[1], Valeria Vitelli[1], Thomas Fleischer[2]

[1] *Oslo Centre for Biostatistics and Epidemiology, Department of Biostatistics, University of Oslo, Norway*
[2] *Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital, Norway*

Variable selection is crucial in high-dimensional omics-based analyses since it is biologically reasonable to assume only a subset of non-noisy features contributes to the data structures. However, the task is particularly hard in an unsupervised setting, and a priori ad hoc variable selection is still a very frequent approach, despite the evident drawbacks and lack of reproducibility. We propose a Bayesian variable selection approach for rank-based transcriptomic analysis. Making use of data rankings instead of the actual continuous measurements increases the robustness of conclusions when compared to classical statistical methods, and embedding variable selection into the inferential tasks allows complete reproducibility. Specifically, we develop a novel extension of the Bayesian Mallows model for variable selection that allows for a full probabilistic analysis, leading to coherent quantification of uncertainties. Simulation studies demonstrate the versatility and robustness of the proposed method in a variety of scenarios as well as its superiority with respect to several competitors when varying the data dimension or data generating process. We use the novel approach to analyse genome-wide RNAseq gene expression data from ovarian cancer samples: several genes that affect cancer development are correctly detected in a completely unsupervised fashion, showing the usefulness of the method in the context of signature discovery for cancer genomics. Moreover, the possibility to also perform uncertainty quantification plays a key role in the subsequent biological investigation.

Parallel Sessions

S7.2 More than meets the eye: Visualising temporal patterns in time-series single-cell RNA-seq data

Laia Canal Guitart[1], Maren Hackenberg[1], Harald Binder[1]

[1] *Institute of Medical Biometry and Statistics, Medical Center, University of Freiburg, Germany*

In a typical analysis workflow for time-series single-cell RNA-sequencing (scRNA-seq) data, a first step is to reduce dimensionality to visually inspect temporal patterns, as there is no one-to-one correspondence between cells at different timepoints. Here, one implicitly assumes that the 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 neighbor embedding (tSNE), uniform manifold approximation and projection (UMAP) and single-cell variational inference (scVI), a popular deep learning-based approach.
To characterise the approaches in a controlled setting, we create artificial time-series scRNA-seq data by applying a specific dimension reduction (say, tSNE) to static snapshot data, and transforming the low-dimensional representation according to biologically meaningful temporal patterns, e.g., dividing cell clusters during a differentiation process. We use deep learning to generate synthetic time-series data in the original high-dimensional gene expression space, and apply all other dimension reduction techniques to compare how well they reflect dynamics happening in, e.g., tSNE space.
Subsequently, we repeat the process using the other dimension reduction approaches to infer a low-dimensional embedding in which we induce the underlying ground truth pattern. Thus, we are able to characterise the different perspective of each technique on a specific temporal pattern, and assess their sensitivity with respect to the representation used to generate the development and to the temporal pattern itself. Furthermore, we provide an illustration on real time-series scRNA-seq data, showing that the choice of dimension reduction is a crucial one that should be carefully made.
Finally, our results provide directions for designing dimension reduction techniques that explicitly respect time structure, thus leveraging the potential of time-series scRNA-seq data for insights into fundamental developmental processes.

S7.3 Prior distributions for structured covariance matrices

Oludare Samuel Ariyo[1]

[1] *Federal University of Agriculture, Abeokuta, Nigeria*

Prior distributions for covariance matrices are a well-studied topic in Bayesian modeling. The most popular priors, such as inverse Wishart, require a completely unrestricted covariance matrix, which is not satisfied in some structural equation models. In these models, the covariance matrices are «structured»: certain covariances in the matrix are fixed to 0 or constrained to be equal. While some prior distributions exist for this situation, the parameterizations are relatively complicated, making it difficult for researchers to meaningfully specify their priors. For example, previous approaches include transforming the covariance matrix to spherical coordinates or placing prior distributions on partial correlations.
In this project, we explore the «naive» way of placing priors on structured covariance matrices. These involve prior distributions for standard deviations separately from correlations, with a univariate distribution placed on each parameter. While these priors provide interpretable parameterizations and allow for varying degrees of informativeness, they are deceiving because they allow for covariance matrices that are not positive definite. This means that if we only consider positive definite covariance matrices under these priors, the priors are more informative than we might naively expect. We discuss a method for generating only positive definite covariance matrices from these naive prior distributions, study the amount of information actually implied by the priors, and provide results on the resulting MCMC algorithms' calibration.
We seek to answer the question, «what exactly do you get when placing naive priors on covariance matrices?»

S7.4 Identification of prognostic and predictive biomarkers in high-dimensional data with PPLasso

Wencan Zhu[1], Céline Lévy-Leduc[1], Nils Ternès[2]

[1] *AgroParisTech, Paris, France,* [2] *Sanofi R&D*

High-dimensional data, including high-throughput genomic data, are becoming increasingly available as sequencing technology advances. Finding significant genomic factors (biomarkers) for clinical outcomes is fundamental in the medical research. Prognostic biomarkers can help to better understand the disease mechanism, and predictive biomarkers serve as a tool for patient treatment strategy. Previous researches were mainly focused on clinical characteristics, and the use of genomic data in such an area is hardly studied. Therefore, methods are required to simultaneously select prognostic and predictive biomarkers in high dimensional genomic data where biomarkers are highly correlated. We propose a novel approach called PPLasso (Prognostic Predictive Lasso) that can achieve this goal with higher accuracy compared to the traditional Lasso. We consider the ANCOVA type approach that integrates prognostic and predictive effects into one statistical model, and then transform the design matrix to remove the correlations between the biomarkers before applying the generalized Lasso. In a comprehensive numerical evaluation, we show that PPLasso outperforms the traditional Lasso approach on both prognostic and predictive biomarker identification in various scenarios. Our method is implemented in the PPLasso R package, available from the Comprehensive R Archive Network (CRAN).

S7.5 Fast marginal likelihood estimation of group-adaptive elastic net penalties

Mark van de Wiel[1], Mirrelijn van Nee[1]

[1] *Amsterdam University Medical Center, Netherlands*

Elastic net penalization is widely used in clinical high-dimensional prediction and variable selection settings. Auxiliary information on the variables, termed co-data, is often available as groups. Examples are chromosomes or pathways for gene expression data. Group-adaptive elastic net penalization exploits this information to potentially improve performance by estimating group penalties, thereby penalizing important groups of variables less than other groups. Estimating these group penalties is, however, hard due to the high dimension of the data. Existing methods, such as multi-grid cross-validation, are computationally expensive or not generic in the type of response. Here, we present a fast method for estimation of group-adaptive elastic net penalties for generalized linear models. We first derive a low-dimensional representation of the Taylor approximation of the marginal likelihood for group-adaptive ridge penalties, to efficiently estimate these penalties. Then, we show by using asymptotic normality of the linear predictors that this marginal likelihood approximates that of elastic net models. The ridge group penalties are then transformed to elastic net group penalties by using the variance function. The method allows for overlapping groups (e.g. gene pathways) and unpenalized variables, such as known clinical risk factors. Moreover, it is easily extended to other penalties or priors, like spike-and-slab.
We apply the method, termed squeezy, to a cancer genomics application. Here, we are interested in prognostic microRNA biomarkers for treatment response of metastasized colon cancer. As co-data we use eight groups of differential expression levels of those miRNAs between tumor and non-cancerous colon tissue. Here, we hypothesize that tumor-specific miRNAs may also be more prognostic. Indeed, we show this to be the case. Moreover, squeezy uses these co-data groups to substantially improve the elastic net in terms of predictive performance. In addition, the selected markers are more stable across resampled versions of the data, which is important for clinical applications as these often require a limited set of markers. We also provide an extensive comparison with other co-data methods. Here, the predictive performance of squeezy is superior or matching to others, where in the latter case the computational performance of squeezy is much better.

Parallel Sessions

PARALLEL SESSION 9: Meta-analysis

S9.1 A New Visualisation for Component Network Meta-Analysis: The Circle plot

Elnaz Saeedi[1], Alex Sutton[1], Suzanne Freeman[1], Clareece Nevill[1], Nicky Welton[2], Deborah Caldwell[2], Jamie Hartmann-Boyce[3], Jose Ordonez-Mena[3]

[1] University of Leicester, United Kingdom, [2] University of Bristol, United Kingdom, [3] University of Oxford, United Kingdom

BACKGROUND: Component network meta-analysis (CNMA) is a method to combine evidence from randomised controlled trials of complex interventions which can be considered to consist of multiple components. Different interventions may include different combinations of components resulting in a complicated evidence base. Understanding the data structure is important to indicate which (if any) CNMA models are possible to fit. This presentation will describe a new plot for visualising the CNMA data to facilitate understanding of the evidence base and analysis possibilities.
METHODS: Following a review of how published CNMA are presented, a circos-type circular plot was developed and refined iteratively to represent the available evidence base for a CNMA. The exemplar dataset was of psychological preparations prior to surgery including 6 treatment components trialled in 19 distinct combinations across 36 trials (54 pairwise comparisons). The primary outcome was length of stay in hospital. The plot was created in R using the "circlize" package.
RESULTS: The plot presents the comparison of component combinations across all trials and the treatment contrasts they estimate. The circular bar chart around the edge of the plot presents the number of individuals randomised to each trial arm; arm-level covariates could also be presented, or a composite bar chart to include the number of events for binary outcomes. We compare our plot to the visualizations used previously for CNMA and highlight the advantages of the developed plot. This includes identifying which components are uniquely identifiable, leading to an understanding of which pairwise interactions are also estimable. The compact composite circular form is visually appealing and extremely efficient in delivering multiple dimensions of information.
CONCLUSION: Since complex interventions and CNMA are increasingly employed, including in public health settings, developing an informative visualization of the data structure is important to help establish i) whether fitting a CNMA model is feasible and, if so; ii) the modelling strategy to be employed. Development of a bespoke software solution to allow multi-arm trials to be accurately represented is currently underway.

S9.2 The usual method of estimating weights for studies in meta-analyses is biased

Stephen Walter[1], Narayanaswamy Balakrishnan[1]

[1] McMaster University, Hamilton ON, Canada

BACKGROUND: In meta-analyses, it is standard practice to weight the component studies by the inverse variance of their estimated treatment effect. The rationale is to assign greater weight to studies with more precise estimates, and to optimise precision of the overall treatment effect. The usual method (apparently universally applied) to obtain study weights is by simply taking the reciprocal of the estimated variance. Unfortunately, this naïve approach gives a biased estimate of the inverse variance, and gives unduly large weight to smaller studies. This bias then also affects the overall estimated treatment effect and its standard error.
OBJECTIVES: To examine the degree of bias in the usual method of calculating study weights, and to develop a suitable bias correction.
METHODS: We first considered the case where the variances are assumed equal in the two treatment groups of a study. By considering the underlying distribution of the treatment effect variance, we established an exact analytic expression for the bias in the study weight. In a second case, where the group-specific variances are assumed to be different, we obtained an approximation for the bias. Bias-corrected study weights can then be used to execute the meta-analysis correctly.
RESULTS: We show that the study weights using the naïve method are always overestimated, particularly in small studies. The variance of the estimated treatment effect is underestimated, possibly substantially. Our bias-correction yields unbiasedness for the estimated treatment effect, which also has minimum variance.We illustrate the bias numerically for typical scenarios, and show examples of how it affects actual meta-analyses.
LIMITATIONS: We assumed that the data are normally distributed. However, this is not a major limitation, because the mean response will be approximately normally distributed in most samples. We have also now developed analogous bias corrections for binary outcome data.
CONCLUSIONS: The standard but biased practice of simply taking the reciprocal of a study variance to establish its weight in a meta-analysis should be abandoned. Our simple correction to eliminate the bias is easily achieved, and validity in estimating the treatment effect is restored.

S9.3 Ranking treatments on multiple outcomes and trade-off between benefit and harms

Virginia Chiocchia[1], Johannes Schneider-Thoma[2], Andrea Cipriani[3], Toshiaki A. Furukawa[4], Stefan Leucht[2], Georgia Salanti[1]

[1] Institute of Social and Preventive Medicine (ISPM), University of Bern, Switzerland, [2] Department of Psychiatry and Psychotherapy, School of Medicine, Technical University of Munich, Germany, [3] University of Oxford, United Kingdom, [4] Departments of Health Promotion and Human Behavior and of Clinical Epidemiology, Kyoto University, Japan

When multiple treatments are available for a specific condition, network meta-analysis can identify which one is preferable by estimating relative treatment effects between interventions for outcomes of interest. From these relative treatment effects, ranking metrics can be calculated to produce treatment hierarchies that rank the interventions from most to least preferable. Among the most reported rankings we find the probability of producing the best value, the SUCRA or P-score. A limitation of these ranking metrics is that they consider only one outcome of interest so often two or more treatment rankings are reported to present the performance of the interventions, especially when both harms and benefits are important. Spie charts are a new approach to measure and visualise the performance of each treatment on multiple outcomes. A spie chart comprises sectors representing outcomes, with their angle reflecting the importance of each outcome. The total area within a spie chart is a metric that can be used to rank the treatments. We extend this approach by combining the area inside the spie chart for efficacy outcomes and the area inside the spie chart for safety outcomes with the concept of net benefit to allow a trade-off between benefit and harms. In our case, efficacy and safety are measured on a probability scale and their trade-off is controlled by the "opportunity cost" threshold equal to $\lambda=\nu u$, where u can be interpreted as how much we are willing to compromise in terms of safety for an increase in efficacy. Then, the value for the new metric can be seen as the difference in terms of probabilities that a treatment is the most efficacious for a specific threshold (in terms of side effects). We present this extended ranking metric, and we illustrate it using two published networks in the field of psychiatry.

S9.4 Using cutting-edge methodology to maximise the value of Individual Participant Data meta-analysis

David Fisher[1], Peter Godolphin[1], Jayne Tierney[1], Mahesh Parmar[1], Ian White[1]

[1] MRC Clinical Trials Unit at UCL, United Kingdom

BACKGROUND: While a major advantage of individual participant data meta-analysis (IPDMA) is the ability to employ a range of analytical approaches, often this is not fully utilised. For example, IPDMA of time-to-event outcomes may not assess proportionality of hazards or consider absolute effects. With participant-level effect modifiers, issues such as power, prognosis, covariate adjustment or missing data may not be considered. In a recent collaborative IPDMA in advanced prostate cancer, we applied a range of cutting-edge approaches to maximize the reliability and clinical utility of results.
METHODS: A full protocol (CRD42019140591) & analysis plan were developed a priori. We obtained comprehensive IPD from three large randomised trials adding docetaxel to standard care. Flexible parametric modelling and standardisation were used to investigate proportionality of hazards and to estimate 5-year absolute survival differences. Analysis of pre-specified participant subgroup interactions used the progression-free survival (PFS) outcome to maximize power, adjusting for a core set of baseline covariates (missing values imputed), and estimated using a within-trial framework to avoid aggregation bias. One-stage modelling was used to analyse multiple interactions simultaneously, and test for higher-order interactions.
RESULTS: We obtained IPD for 2261 men, and confirmed a clear benefit of docetaxel on overall survival and PFS on both relative and absolute scales. Evidence of non-proportionality of hazards was detected & modelled. Of 13 a priori selected covariates, 3 were found to interact with treatment at 10% significance: clinical T stage, Volume [of metastases], and Timing [of metastatic disease]. Simultaneous interaction analysis suggested the effect of Timing was accounted for by the others; however all were correlated. Volume and cT stage together identified a patient subgroup where docetaxel appeared most efficacious; Volume and Timing identified a subgroup where it appeared not to be efficacious.
CONCLUSIONS: We present an innovative demonstration of the estimation of multiple simultaneous within-study covariate interactions in IPDMA, whilst tackling non-proportional hazards via flexible parametric modelling, within a pre-specified protocol. Novel methodology was combined with pragmatic interpretation to guide clinicians in refining treatment approaches.

Parallel Sessions

S9.5 A Simulation Study Comparing Methods for Meta-Analysis of Time-to-Event Outcomes

Theodosia Salika[1], Becky Turner[2], David Fisher[2], Jayne Tierney[2], Ian White[2]

[1] University College London, United Kingdom, [2] MRC Clinical Trials Unit at UCL, United Kingdom

BACKGROUND: Systematic reviews and meta-analysis of time-to-event outcomes are frequently published within the Cochrane Database of Systematic Reviews (CDSR); however, these outcomes are reported differently across meta-analyses. They can be analysed on the hazard ratio (HR) scale or dichotomized and analysed as binary outcomes using effect measures such as odds ratios (OR). If analysed as binary, results on the HR scale may be obtained via the complementary log-log (clog-log) link. A previous meta-epidemiological study identified that dichotomising time-to-event outcomes may be adequate for low, but not for high, event probabilities. We performed a comprehensive simulation study to examine which meta-analysis characteristics affect differences between results obtained on the HR or OR scales.
METHODS: We created 28 simulation scenarios defined by: number of trials per meta-analysis, trial sample size, log HR, between-study variability, follow-up time, and percentage random and fixed censoring. We compared "gold standard" approaches to analysis on the HR scale (Cox and log-rank method) with analysis as binary using either a logit link on the OR scale or a clog-log link on the HR scale.
RESULTS: Analysing time-to-event data as binary using the logit link adversely affected bias, coverage and relative precision in many simulation scenarios. However, in the presence of low percentage of random censoring (probability<0.25)) and low event probability (probability<0.3), analysing the data as binary on the OR scale would be acceptable, if information required to obtain a HR estimate together with uncertainty is absent. Analysing the data as binary using the clog-log link consistently produced more bias, low coverage and low power. Between-study heterogeneity and study sample size did not affect the levels of bias.
CONCLUSIONS: If a HR estimate and variance cannot be obtained for each trial, a meta-analysis using the OR scale (and the logit link) could be an acceptable alternative approach. However, we would recommend performing such meta-analyses only in situations where the event probability and percentage random censoring per trial are low. The clog-log link is an unsuitable approach for analysing the data as binary on a HR scale.

PARALLEL SESSION 10: Ageing

S10.1 4-step longitudinal analysis of latent traits derived from measurement scales in chronic diseases

Tiphaine Saulnier[1], Cécile Proust-Lima[2], Alexandra Foubert-Samier[3], Viviane Philipps[1]

[1] University of Bordeaux, INSERM U1219, Bordeaux Population Health Centre, France, [2] Inserm BPH U1219, Université de Bordeaux, France, [3] University Hospital of Bordeaux, French Reference Centre for MSA, France

INTRODUCTION: There is a growing interest in subjective latent traits, such as quality of life (QoL), to describe progression of chronic diseases. Indeed, as centered on patients'feelings, they provide crucial information for delivering better support. The longitudinal analysis of latent traits requires special attention: (i) it needs to adapt to the definition of the trait as one or multiple constructs underlying the items of a measurement scale as done with the Item Response Theory (IRT); (ii) it needs to account for the possible association with clinical events (e.g. death) as done in the joint modeling (JM) framework; (iii) it needs to be studied in relation to the clinical progression. We propose a 4-step longitudinal analysis to address this issue.
METHODS: The 4 steps are: (1) identify all the underlying dimensions measured by the scale, through factorial analysis; (2) assess each dimension trajectory, using a joint latent process model (JLPM) which models the trajectory of a latent trait measured by ordinal items using a longitudinal IRT model, and simultaneously models the risk of event according to the latent trait. (3) determine the sequence of items' degradation and confront it to disease progression by projecting objective markers along the sequence; (4) quantify the item-Fisher information at different disease stages.
APPLICATION: We applied this strategy to describe QoL changes in Multiple-System Atrophy (MSA), a rare neurodegenerative disease, using data from the French MSA cohort (>600 patients). Repeated measures of the 40 ordinal items from the MSA-specific QoL scale were analyzed revealing 4 dimensions (motor, non-motor, emotional, nutrition) that progressively deteriorated over the course of the disease with faster changes for the motor sphere and modulations of the impairment according to sex, age, and MSA subtype. Most informative items were listed informing on the most sensitive manifestations during clinical progression.
CONCLUSION: The proposed 4-step strategy offers a complete approach for the longitudinal analysis of measurement scales in chronic diseases. By combining IRT along with JM, it addresses the statistical issues due to both the measurement scales and the occurrence of clinical endpoints. The use of Fisher information and clinical progression projections also offers potential for item selection and patient stratification.

S10.2 Biological age cannot be estimated with cross-sectional data

Marije Sluiskes[1], Mar Rodríguez Girondo[1], Hein Putter[1], Jelle Goeman[1]

[1] Leiden University Medical Center, Netherlands

We all age, but the rate at which we do so varies between individuals. The idea behind the concept of 'biological age' is that it better captures one's true underlying global physiological state than chronological age does. Biological age can be captured or tracked by so-called 'biomarkers of ageing', which can be either low- or high-dimensional.
Biological age is latent, which complicates estimation and validation of resulting predictions. Often, biological age estimates are obtained using cross-sectional data and methods, where (candidate) biomarkers are measured at a single point in time and no additional data is available. The differences between predicted age and chronological age are then taken to be informative of one's individual ageing rate.
Even though cross-sectional approaches are widely used, it is unclear to what extent they are of added value in the quest for a reliable and robust biological age estimator. I will show that cross-sectional approaches are superfluous and possibly misleading, as they suffer from two fundamental limitations: 1) they provide no evidence for or against candidate biomarkers as true predictors of biological age and 2) it is impossible to verify whether differences between chronological age and estimated biological age indeed reflect differences in the individual rate of ageing. These two critical limitations hold for any statistical method based on cross-sectional data. I will illustrate this through simulation results, and will conclude by giving some suggestions for possible alternative methods, based on time-to-event data.

S10.3 Multimorbidity pattern and risk of dementia: an 11-year follow-up study using the UK Biobank cohort

Mizanur Khondoker[1], Lee Shepstone[1]

[1] University of East Anglia, Norwich, United Kingdom

BACKGROUND: People with dementia aged over 65 years, on average, have an additional 4.6 chronic conditions. Previous studies have linked some chronic illnesses (such as diabetes, depression, arthritis, hearing impairment) as independent risk factors of dementia, however, little is known how combinations or clusters of potentially interacting chronic conditions may influence the risk of developing dementia.
OBJECTIVES: To identify multimorbidity patterns and investigate their associations with the risk of developing dementia.
METHODS: We used latent class analysis (LCA) on the multivariate binary indicators of 27 chronic conditions identified via linked NHS hospital admission records and death registers of the 447,888 UK Biobank participants. Using a retrospective longitudinal cohort design with a median follow-up duration of 11.3 years, we identified 5,139 incident cases of all-cause dementia (ACD) during the observation period. The outcome was represented as time-to-ACD with participants not developing dementia during the observation window and those who were lost to follow-up or died before developing dementia being considered as censored. The effects of multimorbidity clusters accounting for relevant covariates on the risk of developing dementia were investigated using Cox regression model.
RESULTS: We identified five clinically interesting multimorbidity clusters representing Neurosensory, Mental health (psychiatric), Cardiometabolic, Inflammatory (or autoimmune), and Cancer-related pathophysiology respectively. Compared to people without multimorbidity, membership to multimorbidity clusters dominated by Neurosensory (HR=2.43, p<0.001, 95% CI:1.90 to 2.92), Mental health (HR=2.26, p<0.001, 2.00 to 2.55), and Cardiometabolic conditions (HR=2.23, p<0.001, 95%CI:2.06 to 2.42) showed the highest risk of developing dementia. Contrary to the suggested hypothesis in the recent literature, neither C-reactive protein (CRP) nor APOE genotype was found to moderate the effects of multimorbidity on the risk of developing dementia.
CONCLUSIONS: Early identification of older adults at higher risk of accumulating multimorbidity of specific pathophysiology, and tailored interventions to prevent or delay the onset of such multimorbidity may help prevention of dementia.

S10.4 Expected life years compared to the general population

Maja Pohar Perme[1], Damjan Manevski[1], Per Kragh Andersen[2], Nina Ružić Gorenjec[1]

[1] Institute for Biostatistics and Medical Informatics, Faculty of Medicine, University of Ljubljana, Slovenia, [2] Section of Biostatistics, Department of Public Health, University of Copenhagen, Denmark

For cohorts with long-term follow-up, the number of years lost due to a certain disease yields a measure with a simple and appealing interpretation. Recently, an overview of the methodology used for this goal has been published and two measures have been proposed. In this work, we consider a third option, that may be useful in settings in which the other two are inappropriate.
In all three measures, the survival of the given data set is compared to the expected survival in the general population which is calculated using external mortality tables. In this talk, we will analyze the differences between the three measures, their assumptions, interpretation and the corresponding estimators. The first measure is defined in a competing risks setting and assumes an excess hazard compared to the population, while the other two measures also allow estimation for groups that live better than the general population. In this case, the observed survival of the patients is compared to that in the population, the starting point of this comparison depends on whether the entry into the study is a hazard changing event (e.g. disease diagnosis or the age at which the inclusion criteria were met). The differences between the methods will be also illustrated with examples.

S10.5 Recoverability and estimation of causal effects under typical multivariable missingness mechanisms

Jiaxin Zhang[1,2], Ghazaleh Dashti[2], John Carlin[1,2], Katherine Lee[2], Margarita Moreno-Betancur[1,2]

[1] University of Melbourne, Parkville VIC, Australia , [2] Murdoch Children's Research Institute, Parkville VIC, Australia

In the context of missing data, the average causal effect (ACE) is said to be recoverable if it can be consistently estimated from the available data using an appropriate procedure. Whether this is the case depends on assumptions about the missingness mechanism, which can be specified by adding variable-specific missingness indicators to causal diagrams – so-called missingness directed acyclic graphs (m-DAGs). Previous research defined ten canonical m-DAGs, which capture typical multivariable missingness mechanisms in epidemiological studies, and determined the recoverability of the ACE in the absence of effect modification, but not in the more general setting. In this work, we first determined the recoverability of the ACE in the general setting in simplified m-DAGs, where there were no unmeasured common causes of missingness indicators. Then, to study the possibility of approximately unbiased estimation of the ACE in more realistic settings, we designed a simulation study to evaluate the bias of seven missing data methods when estimating the ACE via correctly specified g-computation: complete-case analysis, simple multiple imputation (MI) without interactions, MI with interactions between exposure and different sets of variables (only outcome, only incomplete confounders, and all confounders), MI where imputation and analysis models were approximately compatible, and MI using the substantive model compatible approach. We generated data based on an example from the Victorian Adolescent Health Cohort Study (VAHCS), where interest was in estimating the ACE of adolescent cannabis use on mental health in young adulthood among females. We evaluated low and moderate exposure prevalence scenarios over four outcome generation models. Missing data were generated according to ten canonical m-DAGs, exploring five scenarios in terms of interaction terms in missingness models. The simulation results show that approximately unbiased estimation is possible with appropriate approaches, even in some settings where the ACE is theoretically non-recoverable. However, approaches incorporating external information (e.g., delta-adjustment methods) are still necessary under some missingness mechanisms. We illustrate the methods in the VAHCS data.

PARALLEL SESSION 11: Communication of survival methods

S11.1 Different multi-state structures when studying antidepressive medication in women with breast cancer

Nikolaos Skourlis[1], Paul Lambert[2], Michael Crowther[1,3], Therese Andersson[1]

[1] Karolinska Institute, Solna, Sweden, [2] University of Leicester, United Kingdom, [3] Red Door Analytics, Stockholm, Sweden

Multi-state models allow the study of complex disease pathways, allowing the formulation and exploration of clinically meaningful research questions. The diagnosis of breast cancer is correlated with psychological distress and the use of anti-depressive medication. In this setting, different research questions may be of interest, such as the probability of ever being prescribed anti-depressive medication, the probability of being in an anti-depressive medication cycle or a post- medication period as well as the length of stay in these periods. The probability of experiencing a second medication cycle after experiencing the first one is also of interest. Different multi-state structures may correspond to different research questions, with each structure having its own traits, advantages, interpretations, and limitations. There are a number of choices in terms of the definition of an anti-depressive medication cycle, the choice of time-scale for different transitions, how to borrow information across transitions, and how to relax the proportional hazards assumption. The aim of this study is to provide insight into the use, interpretation and comparison of different multi-state structures using the Swedish prescribed drug register to study the (re-) occurrence of antidepressive medication cycles and post medication periods for women diagnosed with invasive breast cancer and healthy matched controls from the Swedish population. We also investigate the impact of the different choices mentioned above. The conclusions drawn serve a better understanding of the application of multi-state models in epidemiology.

S11.2 Fair comparison of cause-specific and relative survival by accounting for informative censoring

Molly Wells[1], Paul Lambert[1,2], Mark Rutherford[1]

[1] University of Leicester, United Kingdom, [2] Karolinska Institute, Solna, Sweden

Net survival is commonly used to estimate cancer survival to obtain the probability of surviving cancer in the absence of any other causes of death. Cause-specific survival and relative survival are two methods that are used to estimate net survival. Cause-specific survival only considers deaths from the cancer as the event of interest with deaths from other causes censored. Relative survival compares the observed survival in a cohort of cancer patients relative to that expected in the general population without the cancer of interest. A key advantage of using relative survival is that it does not require cause of death information, unlike cause-specific survival, which can cause problems if there are misclassifications of deaths on death certificates.
Informative censoring is a problem for both methods when estimating net survival. For relative survival, the non-parametric Pohar-Perme estimator is recommended as it is less biased than traditional estimators. To combat the impact of informative censoring, the Pohar-Perme estimator incorporates time-dependent weights giving greater weight to individuals with a higher risk of dying from other causes. However, in cause-specific survival, it is very rare for the informative censoring to be accounted for. Therefore, in the majority of comparisons of relative survival and cause-specific survival, some of the observed differences could be due to accounting or not for informative censoring.We have investigated the use of weighted Kaplan-Meier estimates in cause-specific survival to overcome the impact of informative censoring. The weights for the Kaplan-Meier estimates were obtained using model-based approaches (including using age as the time scale) and using lifetables. We conducted a simulation study that demonstrates better performance for the weighted approach.
We found that using weighted Kaplan-Meier estimates in cause-specific survival analysis provides a better estimate of marginal net survival. In comparisons of cause-specific and relative survival, it is important to compare "like-with-like" and, for fair comparison of the methods, the weighted approach should be considered for cause-specific survival analysis.

S11.3 Cox regression with linked data

Thanh Huan Vo[1], Guillaume Chauvet[2], André Happe[3], Emmanuel Oger[3], Stéphane Paquelet[1], Valérie Garès[4]

[1] Institute of Research and Technology b<>com, France, [2] ENSAI (IRMAR), Campus de Ker Lann, Bruz, France, [3] EA 7449 REPERES, France, [4] Univ Rennes, INSA Rennes, CNRS, IRMAR - UMR 6625, Rennes, France

Record linkage is increasingly used, especially in medical studies, to combine data from different databases which refer to the same entities. The linked data can bring analysts novel and valuable knowledge which is unable to obtain from a single database. However, linkage errors are usually unavoidable regardless of record linkage methods and ignoring these errors may lead to bias estimates. While different methods have been developed to deal with the linkage errors in the generalized linear model, there is not much interest on Cox regression model though this is one of the most important statistical models in clinical or epidemiological research. In this article, we propose an adjusted estimating equation for secondary Cox regression analysis, where linked data have been prepared by someone else and no information on matching variables are available to the analyst. Through a Monte Carlo simulation study, the proposed method has significantly corrected the bias of the parameter estimates of the Cox model caused by false links. An asymptotically unbiased variance estimator is also proposed. Finally, the proposed method will be applied to a linked database from the Brest stroke registry in France.

S11.4 Evaluating cancer screening programmes using survival analysis

Bor Vratanar[1], Maja Pohar Perme[1]

[1] Institute for Biostatistics and Medical Informatics, Faculty of Medicine, University of Ljubljana, Slovenia

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.

S11.5 Assessing lead time bias due to mammography screening on estimates of loss in life expectancy

Elisavet Syriopoulou[1], Alessandro Gasparini[1], Keith Humphreys[1], Therese Andersson[1]

[1] Karolinska Institute, Solna, Sweden

An increasingly popular measure for summarising cancer prognosis is the loss in life expectancy (LLE), i.e. the reduction in life expectancy following a cancer diagnosis. LLE is often used to assess differences across population groups. However, for screened populations, it is unclear whether parts of the differences could be attributed to lead time bias. Lead time is the extra time added due to early diagnosis, that is, the time from tumour detection through screening to the time that cancer would have been diagnosed symptomatically. It leads to artificially inflated survival estimates even when there are no real survival improvements. Partitioning the effect of screening into real and artificial improvements is challenging and requires knowledge of what would have happened in the absence of screening. We used a simulation-based approach to assess the impact of lead time due to mammography screening on the estimation of LLE in breast cancer patients. A natural history model developed in a Swedish setting was used to simulate the growth of breast cancer tumours and age at symptomatic detection. We generated both time to death from cancer (from a flexible parametric survival model) and time to death due to other causes (using mortality rates from Swedish population lifetables); the minimum time was taken as the time to death. Then, a screening programme similar to current guidelines in Sweden was imposed; different scenarios were considered for screening sensitivity and attendance. To isolate the lead time bias of screening, we assumed that screening does not affect the actual time of death. Estimates of LLE were obtained in the absence and presence of screening, and their difference was used to derive the lead time bias. We found that LLE is affected by lead time bias, with a largest absolute bias of 0.61 years for a high screening sensitivity scenario with perfect screening attendance, corresponding to an underestimation of LLE by 7.5%. Bias was reduced to 0.46 years when allowing for imperfect attendance across screening visits. The results of the analysis suggested that LLE is subject to lead time bias, thus requiring special consideration when interpreting comparisons across population groups.

PARALLEL SESSION 13: COVID-19

S13.1 Estimation and interpretation of vaccine efficacy in COVID-19 randomized clinical trials

Hege Michiels[1], An Vandebosch[2], Stijn Vansteelandt[1]

[1] Ghent University, Belgium, [2] Janssen Pharmaceutica NV

An exceptional effort by the scientific community has led to the development of multiple vaccines against COVID-19. Efficacy estimates for these vaccines have been widely communicated to the general public, but may nonetheless be challenging to compare quantitatively. Indeed, the performed phase 3 trials differ in study design, definition of vaccine efficacy and in how cases arising shortly after vaccination are handled. In this work, we investigate the impact of these choices on the obtained vaccine efficacy estimates, both theoretically and by re-analysing the Janssen and Pfizer COVID-19 trial data using a uniform protocol. We moreover study the causal interpretation that can be assigned to per-protocol analyses typically performed in vaccine trials. Finally, we propose alternative estimands to measure vaccine efficacy in settings with delayed immune response and provide insight into the intrinsic effect of the vaccine after achieving adequate immune response.

S13.2 The impact of nosocomial bacterial co-infections on the mortality of COVID-19 patients

Maiju Pesonen[1], Harry Thorpe[2], Jukka Corander[2], Neil MacAlasdair[3], Gerry Tonkin-Hill[2], Anna Pöntinen[2]

[1] Oslo Centre for Biostatistics and Epidemiology, Norway, [2] University of Oslo, Norway, [3] Wellcome Sanger Institute, United Kingdom

Since the early phase of the Covid-19 pandemic, it has been widely observed that mortality is highest among the elderly, immunocompromised individuals and patients with severe co-morbidities. In regions with high burden of endemic nosocomial pathogens there is an elevated risk of hospital transmission of virulent strains during, for example, ventilator treatment, and it remains unclear how the risk posed by these organisms combines with SARS-COV2 infection for the vulnerable patients. To investigate this we developed a holistic approach to simultaneously identify colonization, infection and transmission of all the major nosocomial bacterial pathogens and applied this to a cohort of 257 hospital in-patients from a large university hospital serving the Lombardy region in Northern Italy which was first hit by the pandemic in Europe in early 2020.
To enable capturing all the relevant pathogens present in the gut, upper airways or lungs of the patients, we used targeted culturing of samples from nasal and rectal swabs, as well as sputum or bronchoalveolar wash. DNA extracted from plate sweeps of the detected growth for a sample was whole genome sequenced at high depth to identify all the relevant pathogenic organisms present in the patients in a simultaneous fashion. Combining this wealth of detailed genomic information with the Covid-19 status of the patients and the relevant metadata from the hospital enabled us to investigate pan-pathogen hospital transmission and to assess the causal role of nosocomial pathogen load for patient mortality.
Our results indicate that Covid-19 infection, any secondary infection, and even presence of any major pathogen in respiratory samples increased the mortality risk of the patients, but secondary infections were not found to have an additive contribution to the mortality risk of Covid-19, nor the survival time. Two bacterial species, Acinetobacter baumannii and Pseudomonas aeruginosa, known for their high rate of antibiotic resistance, were associated with the increased mortality risk.

S13.3 Covid-19: waning immunity and the booster dose effect in Israel

Micha Mandel[1], Yair Goldberg[2], Yinon Bar-On[3], Omri Bodenheimer[4], Laurence Freedman[5], Sharon Alroy-Preis[4], Ron Milo[3], Amit Huppert[5]

[1] Hebrew University of Jerusalem, Israel, [2] Israel Institute of Technology, Haifa, Israel, [3] Weizmann Institute of Science, Rehovot, Israel, [4] Israel Ministry of Health, Jerusalem, Israel, [5] Gertner Institute for Epidemiology & Health Policy, Tel Hashomer, Israel

In December 2020, Israel began a mass vaccination campaign against COVID-19 administering the Pfizer BNT162b2 vaccine. A resurgent COVID-19 outbreak began in mid-June 2021 after which a third (booster) dose of the vaccine was approved. In this talk, I will present the statistical challenges in estimating waning immunity and the relative effectiveness of a booster dose, will describe the statistical methodology used to estimate parameters of interest, and will present the results using Israel's national database.

S13.4 Assessing the impact of elevated population mortality rates due to Covid-19 on relative survival

Rachael Stannard[1], Mark Rutherford[1], Paul Lambert[1]

[1] University of Leicester, United Kingdom

The Covid-19 pandemic has greatly impacted cancer survival and the reporting of such metrics through a number of pathways. Due to the cessation of multiple screening services in the UK during the height of the Covid-19 pandemic there has been a shift in stage at diagnosis, and a number of excess deaths due to cancer as a result of these delayed diagnoses. Similarly, many individuals experienced reduced contact with healthcare services, resulting in further excess deaths due to cancer. In addition to excess deaths due to cancer, there has also been an increase in population mortality rates during 2020 and 2021, with many of these deaths due to Covid-19.
Within the net survival framework, relative survival is a highly prevalent survival metric and its estimation relies on population mortality rates given in national lifetables. Relative survival estimates are only valid if these mortality rates are derived from a population which reflects the survival likely to be seen by the cohort of cancer patients, in the absence of cancer. The Covid-19 pandemic has disproportionately affected cancer patients and hence it is important to assess the need for adjusted lifetables, and how to best make these adjustments. Often our analyses concern data spanning several calendar years and hence the suitability of existing lifetables will require consideration for many years in the future.
We have assessed the potential impact of altered population mortality rates on cancer trends over time by simulation. We considered elevated mortality rates across whole calendar years and also cases where the mortality rate was only increased for a particular few months within a calendar year. We assessed the performance of more granular lifetables, taking monthly rather than yearly mortality rates. We compared each scenario to an unaltered dataset to obtain a range regarding the likely impact on relative survival estimates.
Our findings suggest treating population mortality rates more granularly by month rather than by year will improve the estimation of relative survival for the specific periods most impacted by the pandemic. Further work to assess the impact of shielding practices and in-hospital infections amongst cancer patients is required.

S13.5 Estimation of incubation time in relation to quarantine length: the impact of distributional assumptions

Vera Arntzen[1], Marta Fiocco[1-3], Nils Leitzinger[1], Ronald Geskus[4]

[1] Leiden University, Netherlands, [2] Leiden University Medical Center, Netherlands, [3] Princes Maxima Centrum, Netherlands, [4] Oxford University Clinical Research Unit, United Kingdom

INTRODUCTION: The distribution of incubation time (from infection to symptom onset) is a key quantity in the analysis of infectious diseases. It guides decisions on contact tracing and quarantine policies. Typically, the event time of symptom onset is observed exactly, whereas for the time origin (infection) only an exposure interval is known. Assuming that the risk of infection is constant within that interval, data can be made single interval censored by transforming the time scale. To simplify estimation of incubation time, it is common practice to use parametric distributions. The appropriateness of such distributions remains unclear, especially for the right tail of the distribution which informs quarantine policy. As in the tail there are less observations, its estimate will strongly depend on the assumed parametric distribution. Hence, we hypothesize that a semiparametric approach is more appropriate.
METHODS: We give a detailed account of data structure, likelihood formulation and assumptions usually made in literature, and propose a semiparametric method that relaxes one of the assumptions. We also consider an alternative approach based on renewal processes. We compare all approaches and their sensitivity to the imposed assumptions using simulations.
IMPACT: Accurate estimates of the incubation time distribution is important to optimize quarantine policy. We found that misspecification of the parametric model yields considerable bias, that can be largely overcome by using a semiparametric approach.

PARALLEL SESSION 14: Basket trials

S14.1 Sample size determination in basket trials borrowing information across subsets

Haiyan Zheng[1], Michael Grayling[2], Pavel Mozgunov[3], Thomas Jaki[4], James Wason[2]

[1] University of Cambridge, United Kingdom, [2] Newcastle University, Newcastle upon Tyne, United Kingdom, [3] MRC Biostatistics Unit, United Kingdom, [4] University of Cambridge, United Kingdom

Basket trials are increasingly used for the simultaneous evaluation of a new treatment in various patient subgroups. Eligible patients would share a commonality (e.g., a genetic aberration or clinical symptom), on which the treatment may potentially improve outcomes. A number of sophisticated analysis models, which feature borrowing of information between subgroups, have been proposed for enhanced estimation of the treatment effects. Yet development of methods to choose an appropriate sample size appears to fall behind. A widely implemented approach is to sum up the sample sizes, calculated as if the subtrials are to be carried out as separate studies.
We propose a Bayesian approach to sample size determination in basket trials, where patients are randomly assigned to the experimental treatment or a control within each subtrial. Closed-form sample size formulae are derived to enable borrowing of information between commensurate subtrials. Our approach ensures that each subtrial has a specified chance of correctly deciding whether the new treatment is superior to or not better than the control by some clinically relevant difference. Given fixed levels of pairwise (in)commensurability, the subtrial sample sizes are solved simultaneously. Our solution resembles the frequentist formulation of the problem in yielding comparable sample sizes for circumstances of no borrowing. When borrowing is permitted, a considerably smaller sample size is required. We illustrate the application using data examples based on real trials. A comprehensive simulation study further shows that the proposed approach can maintain the true positive and false positive rates at desired levels. A featured web app will also be shown to the audience during the presentation.

S14.2 Information Borrowing in Basket Trials: A Proposal and Evaluation

Libby Daniells[1], Thomas Jaki[2], Pavel Mozgunov[2], Alun Bedding[3]

[1] Lancaster University, United Kingdom, [2] MRC Biostatistics Unit, United Kingdom, [3] Roche

Recent innovation in trial designs to improve study efficiency has led to the development of basket trials in which a single therapeutic treatment is tested on several patient populations, each of which form baskets. Patients across all baskets share a common genetic marker and as such, an assumption can be made that all patients will have a homogeneous response to treatments. Information borrowing procedures utilize this assumption to draw on information regarding the response in one basket when estimating the response rate in others. This can improve power of estimates, particularly in the presence of small sample sizes. By using methods such as the Bayesian hierarchical model (BHM), Calibrated Bayesian hierarchical model (CBHM), Exchangeability-nonexchangeability (EXNEX) model and a Bayesian model averaging (BMA) procedure, notable improvements in power can be achieved. However, this can come at a large cost of inflated error rates, bringing into question the validity of trial conclusions.
We review and compare the performance of the methods mentioned above, whilst also proposing a modification to the EXNEX model (mEXNEX). The standard EXNEX model fixes the borrowing probability prior to the trial; our proposed modification uses a data-driven approach to set these probabilities based on the homogeneity of the response data, measured through Hellinger distances. Through simulation, we show that in the presence of a basket with a heterogeneous response, unlike the other methods discussed, this model can control type I error rates to a nominal level whilst yielding improved power. Under different data scenarios, we also show that this method has the potential to either improve over the standard EXNEX model or perform similarly.

S14.3 Bayesian information sharing methods for a longitudinal basket trial

Lou Whitehead[1]

[1] Newcastle University, United Kingdom

Basket trials are a novel trial approach whereby a single drug or intervention is investigated in multiple subgroups. Primarily, basket trials have been undertaken in Phase II oncology trials, whereby patients are recruited by a common biomarker or genetic mutation at which the treatment is aimed, and then stratified into baskets based on the location of their cancer. Basket trials offer the opportunity to share information between baskets, thereby potentially increasing the power to detect a positive basketwise treatment effect. Basket trials can offer multiple advantages over running a series of separate trials, including reduced sample sizes, increased efficiency, and reduced costs.
Basket trials would be an ideal design for other disease areas which share similar features, for example, chronic ageing related diseases. However, trials in this area frequently have longitudinal outcomes. We therefore extended a recently developed method by Zheng and Wason (2020) for information sharing within a randomised basket trial, to allow for a longitudinal design.
Three other common methods for basket trial analysis were similarly extended: independent (stratified) analysis, Bayesian hierarchical modelling, and the EXNEX method proposed by Neuenschwander et al. (2015). Methods were compared via a simulation study in terms of bias, MSE, type I error rates, and power. We also illustrated the methods on a real dataset taken from the BiCARB trial, which investigated the clinical and cost-effectiveness of oral sodium bicarbonate therapy for older patients with chronic kidney disease.
Results indicate that each method is suited well to different situations. The Zheng and Wason method may be more suitable to a Phase III confirmatory trial, as it relies on fewer assumptions than others in order to share information between baskets. However, all methods that share information improve the power and lower the type I error (in general) over independent analysis.
We found that Bayesian information sharing methods offer significant improvements over independent basketwise analysis in all scenarios. Furthermore, repeated measurements over time offer increased power over a cross-sectional design. Further work is needed to calibrate the Zheng and Wason method for optimal power and error rate control.

S14.4 Bayesian modelling strategies for borrowing of information in randomised basket trials

Luke Ouma[1], James Wason[1], Michael Grayling[1], Haiyan Zheng[2]

[1] Newcastle University, Newcastle upon Tyne, United Kingdom, [2] University of Cambridge, United Kingdom

Basket trials are an innovative precision medicine clinical trial design evaluating a single targeted therapy across multiple diseases that share a common characteristic. Several precision medicine trials of this type are designed as single-arm trials and are now common in early-phase oncology settings, for which several Bayesian methods permitting information sharing across subtrials have been proposed. With the increasing need for randomised evidence in precision medicine, randomised basket trials have gained popularity. Here, the advantages of borrowing information using Bayesian methods could be exploited in two ways; considering the commensurability of either the treatment effects or the outcomes specific to each of the treatment groups between the subtrials. In this work we propose an approach to borrowing over the subtrial groupwise responses ('treatment response borrowing', TRB) based on distributional discrepancy. We contrast the performance of TRB to the widely adopted approach for borrowing over the subtrial treatment effects ('treatment effect borrowing', TEB). Simulation results demonstrate that both modelling strategies provide substantial gains over an approach with no borrowing. TRB outperforms TEB especially when subtrial sample sizes are small on all operational characteristics, while the latter has considerable gains in performance over TRB when subtrial sample sizes are large, or the treatment effects and groupwise mean responses are noticeably heterogeneous across subtrials. Further, we observe that TRB, and TEB can potentially lead to different conclusions in the analysis of real data. Our findings suggest that TRB is preferable in some trial settings as it confers benefit over TEB.

Parallel Sessions

S14.5 Using Empirical Bayes Power Priors for Designing Basket Trials

Lukas Baumann[1], Johannes Krisam[1], Meinhard Kieser[1]

[1] *Institute of Medical Biometry, University of Heidelberg, Germany*

Basket trials are used to test a new treatment or drug in several subgroups. They are currently mostly applied in oncology, where the subgroups usually comprise patients with different locations of the primary tumor, but a common biomarker. Most basket trials are uncontrolled phase II studies, where a binary endpoint such as tumor response is analyzed. Naïve approaches for the analysis of such trials are to either analyze each subgroup individually, or to pool the data of all subgroups before the analysis is conducted. To address the limitations of these approaches, most of the recently proposed designs utilize Bayesian tools to partly share the information between baskets depending on the similarity.
A promising 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 posterior-parameters of the subgroups, where the weights are derived from a similarity measure of the individual posterior distributions.
We show that this design is closely related to the approach of power priors that was proposed for incorporating historical data, specifically to methods using empirical Bayes techniques. Using this connection, we generalize Fujikawa et al.'s design and discuss other power prior methods that can be used for the analysis of basket trials. The designs based on Empirical Bayes power priors are computationally cheap, so that exact computation of operating characteristics such as the type 1 error rate, power and expected sample size is feasible for a moderate number of subgroups. We compare the performance of the designs to other competitive basket trial designs.

REFERENCES: Fujikawa, K., Teramukai, S., Yokota, I., & Daimon, T. (2020). A Bayesian basket trial design that borrows information across strata based on the similarity between the posterior distributions of the response probability. Biometrical Journal, 62(2), 330-338.

PARALLEL SESSION 15: Machine learning for health applications

S15.1 Uncertainty measures of survival predictions with neural networks applied to molecular data

Elvire Roblin[1], Stefan Michiels[2], Paul-Henry Cournède[3]

[1] *Inserm U1018 - MICS CentraleSupelec, France,* [2] *Institut Gustave Roussy / Inserm U1018, France,* [3] *CentraleSupelec, France*

Machine learning models are increasingly used in various domains, especially in healthcare. For instance, multi-layer perceptrons (MLP) models for time-to-event data can be used to predict patient survival probabilities. However, these pointwise estimations are rarely accompanied by uncertainty measures.
This work aims at building confidence intervals at the patient level to quantify the degree of certainty in the model's survival predictions using ensemble methods of neural networks.
We compared existing ensemble methods: Bootstrap, MC-Dropout (Gal et al., 2016), Deep Ensemble (Lakshlinarayanan et al., 2016), and Fixed Bernoulli Mask (Mancini et al., 2020). We applied these methods on three MLP-based survival predictors: CoxCC, CoxTime (Kvamme et al., 2019), and DeepHit (Lee et al., 2018). Synthetic data were simulated, enabling us to estimate a coverage rate of the estimated confidence intervals. We drew 100 simulation datasets of 4,000 individuals each (split into 2,000 for the training set and 2,000 for the test set) with a censoring rate of 20% using the inverse cumulative distribution method (Bender et al., 2005). These models were also applied on 2 datasets: the METABRIC breast cancer data, including 1,960 patients, 6 clinicopathological covariates, the expression of 1,000 genes, and lung cancer data (LCE) consisting of 4,120 patients, three clinical variables, and 1,000 genes.
On the simulation datasets, CoxTime obtained a mean value of the 100 coverage rates at the median survival time of 0.713 (±0.16 SE) using Bootstrap. It was 0.714 (±0.147) for Deep Ensemble, 0.796 (±0.132) for MC-Dropout, 0.916 (±0.041) for Bernoulli Mask. The highest coverage for CoxCC was achieved at 0.754 (±0.135) with MC-Dropout; for DeepHit, 0.719 (±0.242) with Bootstrap. Overall, the best coverage rate was achieved with Bernoulli Mask associated with CoxTime, which was applied to the data sets. In METABRIC, C-indices of the clinical only and clinic-genomic model were given by 0.740 (±0.008 SE estimated with Bernoulli Mask) and 0.651 (±0.038); while in LCE, they were 0.641 (±0.005) and 0.678 (±0.016). Applying this method on a particular METABRIC patient, the mean survival probability at 5 years and confidence interval for a patient with a censored survival time of 8.5 years was 0.763 [0.419, 0.937].

S15.2 Subgroup discovery with survival decision trees: detection of early conversion in Alzheimer's stages

Marie Laurent[1], Martin Prodel[1], Alexandre Vainchtock[1], Alexandre Civet[2], Antoine Chevrette[3], Beatrice Rousselle[2], Claire Moutet[4],

Virginie Dauphinot[4], Pierre Krolak-Salmon[4], [1] *HEVA, Lyon, France,* [2] *Roche France,* [3] *Roche Global,* [4] *Clinical and Research Memory Centre of Lyon, Lyon University Hospital, France*

INTRODUCTION: Alzheimer's disease (AD) is a neurodegenerative disease resulting in cognitive loss that progress towards functional impairment. The objective of this study was to identify, among patients at a mild neurocognitive disorders (NCD) stage, subgroups of patients converting prematurely [or late] to the major NCD stage (dementia).
METHODS: Data of patients with mild NCD due to clinical AD between 2014 and 2019 were extracted from the MEMORA study, a prospective cohort collected in memory centers of the Clinical and Memory Research Center of Lyon (France). Data include patients' etiological diagnosis, socio-demographic information, comorbidities, treatments, MMSE scores (Mini Mental State Examination) and hospitalizations. To identify profiles at high risk of early/late conversion to the major NCD stage, we trained several survival decision trees, which are regression trees where the quality of a split is measured by the log-rang splitting rule. The censored time to conversion (TTC) was the target variable and all the other variables were predictors. Each tree was trained on a different subset of predictors to increase subgroup diversity. Subgroups were identified among any node or leaf of the trees for which the median TTC (computed through a Kaplan-Meier estimator) is at least 25% far from the overall median TTC. The best subgroups were selected among subgroups larger than 5% of the cohort.
RESULTS: 1,548 mild NCD patients were included in this analysis, with a median TTC of 24 months. Patients with a MMSE ≤ 23 had an early conversion (TTC=17 months, N=703 patients). Patients with a MMSE ≤ 23 and a cardiopathy converted even earlier (TTC=11m, N=112), so did patients with a MMSE ≤ 23 and not under an antidementia treatment (TTC=15m, N=529). Patients with a MMSE ≥ 28 had a late conversion (TTC=43m, N=172), so did patients aged 85 or less, having at least a lower secondary education level, without hypertension nor depression (TTC=40m, N=205).
CONCLUSION: Combining survival decision trees and the MEMORA database enabled the detection of patient subgroups converting especially early/late to major NCD stage. Comorbidities (depression, hypertension), educational level and antidementia drugs appeared as strong predictors of the TTC. As expected, the MMSE score was the main predictor of the model, which confirmed its importance in the AD assessment.

S15.3 Spatio-Temporal Functional Principal Component Analysis

Sonia Dembowska[1], Jeanne Houwing-Duistermaat[2], Haiyan Liu[1]

[1] *University of Leeds, United Kingdom,* [2] *School of Mathematics, University of Leeds, United Kingdom*

Images are frequently used in clinical settings, but statistical models are still limited. Our work is motivated by experimental data on brain fMRI images captured on 15 subjects whilst performing decision-making tasks. The images were acquired every 2.5 s during the investigation providing each subject with a time series of 3 dimensional images representing their brain activity, each patient also received a clinically-derived risk score at the end. The goal of the study was to associate the images to the clinical outcome. The method used to achieve this goal reduced the temporal dataset to one image by taking the difference of the first three images in one time series. This approach is arbitrary and does not address the full response of a subject over time. We propose a novel method using functional data analysis which decomposes the spatio-temporal images into time-independent principal components and time-dependent score functions.
Our functional principal component analysis (FPCA) method preserves the spatial relationships between pixels in the observed data including the temporal aspect. The components hold the spatial information and are three-dimensional whilst we let the scores depend on time and represent individual variation with respect to the components. The novelty of the method is two-fold: firstly, we create a functional principal component analysis model to deal with three-dimensional images for clinical analysis where space and time are detangled. Secondly, the method of parameter estimation is done such that computational requirements are minimal, as we avoid calculating the covariance matrix used to estimate principal components. This makes our method computationally fast.
Data analysis reveals that our score functions are able to predict the clinical risk with high accuracy, outperforming the previous method that is unable to account for time. In 10 of the tasks performed, the individual score functions have predicted the risk rating of subjects with an average RMSE of 0.788 whilst the average RMSE of the old method is 2.497. Our result shows meaningful insights and better clinical predictions can be obtained when modelling temporal aspects of image data.

S15.4 Causal machine learning and use of sample splitting in settings with high-dimensional confounding

Susan Ellul[1], John Carlin[1], Stijn Vansteelandt[2], Margarita Moreno-Betancur[1]

[1] *University of Melbourne and Murdoch Children's Research Institute, Australia*
[2] *Ghent University, Belgium*

Modern day health research aims to address causal questions where there are multiple potential confounding variables at play with complex relationships between them. Singly robust methods for estimating causal effects, such as g-computation or inverse probability weighting, are problematic in such a setting, with misspecification bias likely. Doubly robust approaches such as Augmented Inverse Probability Weighting (AIPW) and Targeted Maximum Likelihood Estimation (TMLE) enable the use of machine learning (ML) to reduce misspecification bias for point estimates, although it has been found that biased standard errors can result when using these methods with very flexible, complex ML approaches. Sample splitting, whereby outcome and exposure models are fit in subsets(folds) of the data and the causal effect estimator is evaluated on the remaining data, has been proposed to improve the estimation of standard errors in this context. Despite advances in these approaches, there is a lack of guidance for implementing AIPW and TMLE in realistic settings like modern observational studies where high-dimensional confounding is present. Furthermore, there remain unanswered questions regarding the use of sample splitting with these approaches, specifically whether cross-fitting, a recommended form of sample splitting, should be used in all situations, and if so, with what number of folds.
Based on a case study from the Barwon Infant Study, we conducted an extensive simulation study to address these questions in the context of estimating the average causal effect. We generated exposure and outcome from regression models with simple (main effects) and complex (interaction and non-linear terms) specifications across scenarios with varying sample size and degree of high-dimensional confounding, represented by two confounder sets (demographic confounders only, demographic plus numerous continuous metabolite confounders). We evaluated the performance of AIPW and TMLE without cross-fitting, and with cross-fitting with 2, 5 and 10 folds, while also varying the diversity of the Super Learner library. In the case study there was large variability in findings for AIPW and TMLE when applying the methods with and without cross-fitting, and when using different libraries, so our simulation findings will provide important guidance on the application of methods in realistic scenarios.

S15.5 Modeling high-dimensional interaction problems with the pliable lasso

Theophilus Quachie Asenso[1], Manuela Zucknick[2], Hai Hai Zhang[3], Liang Yong[4]

[1] *Oslo Centre for Biostatistics and Epidemiology (OCBE), University of Oslo, Norway,* [2] *University of Oslo, Norway,* [3] *Northwest University, Kirkland WA, United States,* [4] *Macau University of Science and Technology, Taipa, Macau*

Modelling interactions in high-dimensional data is a notoriously difficult problem. Analyzing high-dimensional data with conventional tools is very challenging. One of the reasons is that most existing models cannot easily handle cases with high-dimensional data and many interaction effects among the covariates, and they often make strong assumptions, e.g. strong hierarchy between main and interaction effects. In this work we apply the pliable lasso penalty to estimate interaction effects and extend the existing linear pliable lasso model to other models. We also demonstrate how the pliable lasso can be used to estimate interactions.
In the first part, we study multinomial logistic regression with interaction effects. Our approach involves the implementation of the pliable lasso penalty which allows for estimating the main effects of the covariates X and interaction effects between the covariates and a set modifiers Z. The original log-likelihood from the linear model formulation is transformed into an iteratively reweighted least square problem for multinomial logistic regression with the pliable lasso penalty.
The results from the simulation and real data show the effectiveness of the pliable lasso penalty in multinomial logistic regression, and that it has some good qualities and can perform well in the multi-classification problems, involving interaction effects.
In the second part, we will report results from ongoing work from the implementation of the pliable lasso penalty in modeling and predicting synergistic effects between two drugs in drug combination experiments, using for example the molecular characterization of a cell line with multi-omics data to predict, whether two drugs will act synergistically on that particular cell line. This work involves the extension of the pliable lasso model to high-dimensional multi-response regression.

PARALLEL SESSION 17: Competing risks

S17.1 Estimands for Recurrent and Terminal Events: Methods, Issues and Recommendations

Alexandra Bühler[1]

[1] *University of Waterloo, Waterloo ON, Canada*

Recurrent event data arise commonly in health research: recurrent metastases in advanced breast cancer, recurrent disability in progressive forms of multiple sclerosis, or recurrent myocardial infarctions and strokes in cardiovascular disease. There is therefore growing interest in using recurrent event data for the evaluation of therapeutic interventions and treatments in clinical trials. Complexity of such disease processes, however, raises challenges in the specification of estimands for intended treatment comparisons; in particular when the recurrent event process is terminated by death. Cardiovascular death or non-cardiovascular death, for instance, precludes the occurrence of non-fatal myocardial infarctions or strokes. Some guiding principles for defining estimands in complex life history processes will be presented and comprehensively discussed in the context of recurrent and terminal events. Several methods have been proposed for the analysis of recurrent event processes terminated by death in the setting of a clinical trial, with the semi-parametric rate-based models by Ghosh-Lin and Mao-Lin among the most commonly adopted. Despite the widespread use of these methods, details on characteristics of their corresponding estimands are often lacking. In this talk, numerical results based on large sample theory are provided to evaluate estimands arising from marginal Ghosh-Lin and Mao-Lin analyses in terms of their interpretability, estimability, assumptions and robustness. Data from a metastatic breast cancer trial are used to illustrate these points, and also to discuss the usefulness of a utility-based approach.

S17.2 Separable Effects of Baseline Exposure in Multi-State Models

Niklas Maltzahn[1], Jon Michael Gran[1], Mats Julius Stensrud[2]

[1] *Oslo Centre for Biostatistics and Epidemiology, Norway*
[2] *École Polytechnique Fédérale de Lausanne, Switzerland*

Multi-state outcome processes, which generalize classical time-to-event outcomes, are often studied in medical and epidemiological applications. While recent work on causal inference has clarified the causal interpretation of a range statistical estimands in event history analysis, development in the more general setting of multi-state modelling is still limited. Furthermore, most of the existing results, in particular on counterfactuals, have primarily concerned discrete-time random variables. We present a framework for counterfactual causal inference of marked point processes based on counterfactual probability measures, rather than counterfactual random variables. This framework provides a transparent way of relating counterfactual parameters of interest (in particular counterfactual hazards) to properties of the underlying probability measure. To motivate the use of counterfactual probability measures we argue that it is compatible with Robins's FFRCISTG model, and more generally with the SWIGs framework of Robins and Richardson. Using this framework, we develop new results for the study of separable (mechanistic) causal effects in continuous time multi-state settings, inspired by Robins and Richardson's seminal treatment decomposition idea. In the setting of competing events this idea has resulted in new estimands for component specific effects in the treatment of prostate cancer and the generalization to multi-state processes is likewise important for identification of counterfactual parameters of substantive interest.

Parallel Sessions

S17.3 Externally Validating Clinical Dynamic Prediction Joint Models for Localised Prostate Cancer

Harry Parr[1], Nuria Porta[1], Alison Tree[2], David Dearnaley[1], Emma Hall[1]

[1] *The Institute of Cancer Research, London, United Kingdom*
[2] *The Royal Marsden NHS Foundation Trust, United Kingdom*

BACKGROUND: Prostate Cancer (PCa) is highly prevalent and the 2nd most diagnosed cancer in men worldwide. Most PCa diagnoses are early-stage, where disease is localised and has not spread. Following several randomised clinical trials (RCTs), it is well established that moderately hypofractionated (fewer but larger doses of) external beam radiotherapy (RT) delivered with curative intent, is non-inferior to conventional fractionated treatments and is now a standard-of-care for treating non-metastatic PCa. Clinical dynamic prediction joint models (CDPJMs) have been previously proposed to incorporate longitudinal repeated PSA readings over time to improve predictions of PCa prognosis, in addition to using baseline risk factors. We develop here and externally validate a CDPJM in the context of new treatment pathways with hypofractionated RT.
METHODS: The shared-parameter CDPJM is developed using data from CHHiP (N=3,071) – the largest known phase-III hypofractionation RCT. We jointly model patients' baseline risk factors, and the longitudinal repeated PSA biomarker, to predict time-to-recurrence of PCa. We assess prognostic performance. External validation (EV) is performed from two other large localised PCa RCTs: RADAR & RT01, and are compared with internal validation measures. Various landmark times within clinically relevant prediction windows of interest are considered, to assess the CPDJM's performance and generalisability in other patients and healthcare settings. Miscalibration is inspected between model predictions and observed risk, with EV cohorts used to recalibrate accordingly.
RESULTS: We assessed internal validity using x50-bootstrap samples, showing the model to have robust discrimination and calibration. The internal results are optimal after five-years' worth of longitudinal PSA data. The optimism-corrected performance measures are compared with EV, and further recalibration is explored for EV sub-populations with more locally advanced PCa.
CONCLUSION: External validation shows broadly consistent and comparable results, although recalibration may be necessary for more advanced PCa patients, or even further meta-analysis may be necessary for extended model development. These externally validated results are encouraging, we show the CDPJMs can have clinical utility to predict PCa recurrence in localised patients treated with radiotherapy.

S17.4 Integrating relative survival in multi-state models—a non-parametric approach

Damjan Manevski[1], Hein Putter[2], Maja Pohar Perme[1], Edouard F Bonneville[2], Johannes Schetelig[3], Liesbeth C. de Wreede[2,4]

[1] *Institute for Biostatistics and Medical Informatics, Faculty of Medicine, University of Ljubljana, Slovenia,* [2] *Biomedical Data Sciences, LUMC, Leiden, Netherlands,* [3] *Dresden University Hospital, Germany,* [4] *DKMS, Dresden/Tübingen, Germany*

Multi-state models provide an extension of the usual survival/event-history analysis setting. In the medical domain, multi-state models give the possibility of further investigating intermediate events such as relapse and remission. In this work, a further extension is proposed using relative survival, where mortality due to population causes (i.e. non-disease-related mortality) is evaluated. The objective is to split all mortality in disease and non-disease-related mortality, with and without intermediate events, in datasets where cause of death is not recorded or is uncertain. To this end, population mortality tables are integrated into the estimation process, while using the basic relative survival idea that the overall mortality hazard can be written as a sum of a population and an excess part. Hence, we propose an upgraded non-parametric approach to estimation, where population mortality is taken into account. Precise definitions and suitable estimators are given for both the transition hazards and probabilities. Variance estimating techniques and confidence intervals are introduced and the behaviour of the new method is investigated through simulations. The newly developed methodology is illustrated by the analysis of a cohort of patients followed after an allogeneic hematopoietic stem cell transplantation. The work has been implemented in the R package mstate.

S17.5 Degrees of necessity and of sufficiency for competing risks survival data

Andreas Gleiss[1], Michael Gnant[2], Michael Schemper[1]

[1] *Center for Med. Statistics, Informatics, and Intelligent Systems, Med. University of Vienna, Austria,* [2] *Comprehensive Cancer Center, Medical University of Vienna, Austria*

Recently, the authors have proposed measures to quantify the degree to which the effect of prognostic factors is necessary or sufficient for a dichotomous or survival outcome. A condition or cause is considered necessary for an event, if without the cause the event cannot develop. It is considered sufficient if the event is unavoidable in the presence of the cause. Necessity and sufficiency can be seen as the two faces of causation, and this symmetry is reflected by the suggested measures. They provide an approximate multiplicative decomposition of explained variation (EV).
For these measures, the prognostic factors may be measured on different scales or may be of different types (dichotomous, qualitative, or continuous) but the result always is given on a 0-to-1 scale. Thus, explained variation permits to establish a ranking of the importance of factors, while the degree of necessity (DN) and the degree of sufficiency (DS) provide a more nuanced assessment of this importance.
The occurrence of a competing event either makes it impossible to later observe the event of interest or may substantially alter the interpretation of a possibly later following event of interest. While the original EV, DN and DS measures build on comparisons of unconditional and conditional survival functions, these are now replaced by unconditional and conditional cumulative incidence functions, respectively. The latter are often obtained from fitted Fine & Gray models. For estimation, individuals are down-weighted after experiencing the competing event, as proposed by Fine & Gray. Other characteristics of the original measures remain unchanged, such as the approach taken to obtain consistent estimates in the presence of censoring.
Empirical properties of the suggested measures are investigated in an extensive simulation study. Advantages of the approach are exemplified by re-analyzing an Austrian study of 2020 patients with postmenopausal, receptor-positive breast cancer. While the primary outcome, time to distant metastases, had been experienced by only 14%, the competing event, death without prior distant metastases, had been observed in 19% of the patients. N-stage, Grading and log of tumor size exhibit a moderate DN (0.33 to 0.40), but low DS (0.03 to 0.15).

PARALLEL SESSION 18: Personalized Medicine

S18.1 Resampling Methods to Control the Family Wise Error Rate for Dual Biomarker Threshold Identification

Ben Lanza[1], Deepak Parashar[1], Nigel Stallard[1], Chris Harbron[2]

[1] *University of Warwick - Warwick Medical School, United Kingdom,* [2] *F. Hoffmann - La Roche*

BACKGROUND: Patient subgroups are often defined by continuous biomarker values, alongside a threshold to dichotomise the population into sensitive and non-sensitive. There is increasing evidence to use multiple biomarkers to sufficiently identify sensitive patients for some drug combinations. When assessing candidate subgroups within a clinical trial, the Family Wise Error Rate (FWER) must be controlled. Traditional methods of FWER control are often conservative, as they do not incorporate the dependence structure between tests. Resampling based methods, such as the Romano and Wolf step-down multiple testing procedure, can account for the dependence among tests and give increased power to detect subgroup effects whilst controlling the FWER.
METHODS: We identify optimal dichotomising thresholds for two predictive continuous biomarkers simultaneously, whilst assessing the overall treatment efficacy. A grid search over candidate threshold combinations was conducted to identify the optimal subgroup within a single stage trial, using the Romano and Wolf procedure to control the FWER. A simulation study was carried out, alongside an application to a real-world dataset.
RESULTS: Accurate threshold identification was achieved using a range of candidate threshold grid sizes. Accuracy of threshold identification fell as the treatment magnitude and expected subgroup size decreased; accuracy was also lower when the treatment was broadly effective. The FWER was controlled in null cases for implemented grid sizes; we will present the impact of varying grid size on FWER.
CONCLUSIONS: The use of resampling methods allows for dual biomarker threshold identification alongside the assessment of the overall treatment effect, whilst controlling the FWER.

Parallel Sessions

S18.2 Personalized optimal treatment timing through multi-state modelling and microsimulation

Caterina Gregorio[1], Marta Spreafico[1], Francesca Ieva[1]

[1] *MOX – Department of Mathematics, Politecnico di Milano, Italy*

BACKGROUND: In many clinical settings the only possibly curative treatment can be performed once and its timing needs to be tailored based on patients' characteristics. However, the medical decision on the waiting time before treatment is complex. It depends on the risk of treatment failure and on the risk of disease progression over time. In the absence of a trial, it is of interest to use observational data to study the effect of a timing policy. This problem has been only recently faced in the literature. However, existing methods do not consider the case of effect heterogeneity.
OBJECTIVE: We aim at providing a method to estimate the conditional causal effect of the timing of a treatment on a time-to-event outcome to optimize it in a personalized manner.
METHODS: Data comes from a retrospective cohort study involving 2476 patients diagnosed with myelodysplastic syndromes. In this disease deciding when to perform the bone-marrow transplant is very important because of its success and hence patients' prognosis is believed to be greatly influenced by its timing. Since there are often subpopulations in which the effect of an intervention is not the same, we target the conditional Restricted Mean Survival Time (RMST). We use a multistate approach combined with inverse probability of treatment weighting to estimate the expected outcome after intervention. A semi-Markov multi-state model with states "Alive pre-treatment", "Alive post-treatment" and "Dead" was used to describe patients transitioning between disease and treatment states in a counterfactual framework. Conditions for the presence of effect modification in this setting are obtained. Under certain assumptions, we are able to estimate for different waiting times the stratified RMST. The estimation can be either performed analytically or through microsimulation and uncertainty is assessed through bootstrap.
RESULTS & CONCLUSIONS: We provide a method to identify subgroups of patients for which the hypothesis of homogeneous effect of treatment timing does not hold. We then propose a decision analysis to tailor the optimal timing in these subgroups. The developed methodology gives novel insights with regards to the optimal transplant time in myelodysplastic syndromes, according to different genomic profiles.

S18.3 Methods for estimating personalized treatment recommendations with extensions to survival data

Jennifer Hellier[1], Andrew Pickles[1], Richard Emsley[1]

[1] *King's College London, United Kingdom*

BACKGROUND: Personalized medicine requires predictive markers to identify the treatment that is likely to provide the most benefit for the individual. When these markers are combined into a single moderator index, it can be used to calculate a personalized treatment recommendation (PTR) to explore the benefit of treatment for an individual when compared to other treatment decisions. To date, most methods for identifying PTRs have focused on continuous or binary outcomes.
OBJECTIVE: We will present and contrast two approaches to forming a weighted combination of individual markers and identify PTRs. We will evaluate their effectiveness by quantifying the expected benefit of the PTR compared to alternative treatment strategies. We will extend the methods to survival data using the restricted mean survival time (RMST). We will demonstrate the methods with application to a randomised trial in children with ADHD and intellectual disability (HSEN: the Hyperactivity and Special Educational Needs Study) which compared methylphenidate treatment to placebo.
METHODS: The combined moderators are formulated via two methods; (1) a regression approach, a linear combination of treatment by marker interactions; and (2) Kraemer's approach, where individual markers have weights determined using all pairwise comparisons between the randomised arms. For survival data, the RMST is estimated from integration of smooth survival curves over the maximum study period, and an individual's RMST considered as the outcome. The combined moderators are evaluated using the Stata PTR command.
RESULTS: The regression approach outperforms Kraemer's model in terms of efficiency, although model estimates of the combined moderators and estimated benefit of the PTR are comparable. We show that the distribution of the estimated RMST is suitable for calculation of a PTR and captures the treatment heterogeneity in the data. We show that both approaches form a qualitative combined moderator in the HSEN trial data.
CONCLUSION: The restricted mean survival time has proved to be a suitable endpoint to determine personalized treatment recommendations. Multiple markers should be evaluated and validated as a PTR to suggest optimal treatment allocations.

S18.4 Drugs combinations screening using a Bayesian ranking approach based on dose-response models

Luana Boumendil[1], Raphaël Itzykson[1], Vincent Levy[3], Lucie Biard[1]

[1] *Université Paris Cité, France,* [2] *Université Paris Cité-APHP-Inserm, France,* [3] *Sorbonne Paris Nord-APHP-Inserm, France*

BACKGROUND: Drug combinations have been of increasing interest in recent years for the treatment of complex diseases such as cancer, as it could reduce the risk of drug resistance, aiming simultaneously at different pharmacological disease targets. Moreover, combining drugs may allow tailoring therapy to the patients' mutations. Acute Myeloid Leukemia (AML) exemplifies settings for ex-vivo drug combinations screening. With a poor prognosis, especially in older patients, and a high heterogeneity of genetic profiles, the current «one size fits all» strategy is unsuitable for AML, rendering tailored strategies of prime interest. Provided culture environments, leukemic cells can be sampled from patients, to screen candidate drugs. Identifying potent combinations can be an arduous task since exploring the full dose-response matrix of candidate combinations is costly and sometimes unfeasible, as the quantity of available biological material is limited and may vary across patients.
OBJECTIVES: To develop a rank-based screening approach for drugs combinations in AML based on ex-vivo assessment of dose-responses relationships.
METHODS: A Bayesian 4-parameter log-logistic (FFPL) model was used to estimate dose-response curves and dose-candidate combinations were explored using the cross design method [Malyutina 2019] and ranked according to various pre-defined metrics. A final ranking of combinations was obtained using the Surface Under the Cumulative Ranking curve (SUCRA) method [Salanti 2011] based on the posterior distribution of combinations ranking. We evaluated the operating characteristics of the ranking algorithm in a simulation study, considering various scenarios: number of drugs, of patients, shape of dose-response curves, etc. We applied our proposed method to real data of ex-vivo evaluation of drugs combinations in AML patients.
RESULTS: The Bayesian FFPL model allowed estimating the dose-response curves and, using the cross design, the combinations' performance metrics. The SUCRA method performed well in ranking combinations, even in situations with small sample sizes.
CONCLUSIONS: Although the SUCRA approach has been developed for network meta-analyses, we believe it could be applied to drug screening settings. It could reveal to be a useful tool for prioritizing investigational therapies in high throughput treatment screening trials.

S18.5 Dynamic interventions determined by recurrent events

Matias Janvin[1]

[1] *Ecole Polytechnique Federale de Lausanne, Switzerland*

In studies of medical treatments, individuals often experience posttreatment events that can predict their future treatment response. In this work, we study how to use initial observations of a recurrent outcome to offer updated treatment recommendations. We formulate an estimand for the effect of switching between two treatment arms on future event recurrences, conditional on an initial treatment response on the recurrent outcome. Next, we describe the causal interpretation of the estimand and provide partial identification results, which we use to propose dynamic treatment rules with a decision-theoretic justification. The most conservative of these rules is guaranteed to perform at least as well as the optimal treatment when we restrict such that only one treatment can be given during the whole course of the study. We present simulations to illustrate properties of our proposed method and its relation to previous work. Within-subject dependency of recurrent events has previously been modelled using latent frailty variables. However, commonly used frailty models require strong parametric assumptions, and assume homogeneous treatment effects between individuals on a multiplicative scale. Recently, quantile regression has also gained attention as a potential statistical aid to customized disease management. We clarify the conditions required to interpret functionals of quantile regression coefficients as causal effects of dynamic treatment switching. As an illustrative example, we use our proposed method to estimate non-parametric bounds for the effect of dynamic treatment switching on recurrent adverse events in data from the Systolic Blood Pressure Intervention Trial. Under the bounds obtained, we show that the optimal decision will depend on the decision-maker's risk preferences.

Parallel Sessions

PARALLEL SESSION 19: Clinical Trials

S19.1 Should the two-trial paradigm still be the gold standard in drug assessment?

Stella Jinran Zhan[1], Cornelia Ursula Kunz[2], Nigel Stallard[1]

[1] *University of Warwick, Coventry, United Kingdom,* [2] *Boehringer Ingelheim Pharma GmbH & Co. KG*

Two significant pivotal trials are usually required for new drug approval by a regulatory agency. This standard requirement is known as the two-trial paradigm. However, several authors have questioned why we need exactly two pivotal trials, what statistical error the regulators are trying to protect against, and how it compares with alternative approaches, such as a meta-analysis of several studies. Therefore, it is important to investigate these questions to better understand the regulatory decision-making in the assessment of drugs' effectiveness.
It is common that two identically designed trials are run solely to adhere to the two-trial rule. Previous work showed that combining the data from the two trials into a single trial (one-trial paradigm) would increase the power while ensuring the same level of type I error protection as the two-trial paradigm. However, this is true only under a specific scenario and there is little investigation on the type I error protection in the full parameter space.
In this work, we compare the two paradigms by considering 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. Hence, there is the need for appropriate flexibility around the two-trial paradigm and the appropriate approach should be chosen based on the questions we are interested in.

S19.2 Conditional Drug Approval with the Harmonic Mean Chi-Squared Test

Charlotte Micheloud[1], Manja Deforth[1], Leonhard Held[1], Kit Roes[2]

[1] *Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Switzerland,* [2] *Department of Health Evidence, Section Biostatistics, Radboud University Medical Center, Nijmegen, Netherlands*

In case of an unmet medical need, such as a rare disease or a public health emergency, drug approval might be granted under a conditional license. In this case, the evidence for a treatment effect from the first (pre-market) clinical trial later has to be substantiated in an independent post-market trial or a longer follow-up duration. A recent example is the conditional marketing authorization of the COVID-19 Vaccine AstraZeneca by the European Medical Agency (EMA).
Several methods exist to quantify the overall evidence provided by the two trials. We study the applicability of the recently developed harmonic mean $\chi 2$ -test to the conditional drug approval framework. The proposed approach can be used both for the design of the post-market trial based on the results of the pre-market trial, and the analysis of the combined evidence provided by both trials. In contrast to Fisher's criterion and Stouffer's method, the harmonic mean $\chi 2$ -test always requires a post-market clinical trial. Furthermore, as compared to the two-trials rule, the proposed method tends to require a smaller sample size for the post-market trial, while simultaneously leading to a larger overall power to detect the treatment effect.
For illustration, we apply the harmonic mean $\chi 2$ -test to the drug Fampridine, a treatment for patients with Multiple Sclerosis. The drug received conditional (and eventually final) market approval by the EMA. We revisit this study and show how the proposed method could have led to a sample size reduction in the post-market trial. A simulation study was also conducted to investigate the operating characteristics of the methods in more detail.

S19.3 Frequentist and Bayesian approaches to rescuing disrupted trials

Richard Emsley[1]

[1] *King's College London, United Kingdom*

There is a threat to the validity of clinical trials that were underway before COVID-19 pandemic and that were affected by the pandemic response. Complications may arise from pandemic-related operational challenges such as site closure, travel limitations and interruptions to the supply chain for the investigational product, or from health-related challenges such as COVID-19 infected trial participants. Many studies were paused and recruitment restarted without due consideration of whether all studies should restart, or required changes in sample sizes, potentially leading to huge research waste.
We consider a hypothetical randomised trial to compare second line therapies in Type 2 Diabetes. The trial was paused after 60% of the participants had been recruited, and follow-up is completed on all these participants. Before deciding whether or not to restart the study, the trial team should ask: Should we restart recruitment, and are any changes required to the planned sample size? Should we analyse the data that has been collected and draw appropriate conclusions? Can we change the design to an adaptive or group sequential design and incorporate an unplanned interim analysis? How will we account for changes in the patient population, the intervention delivery or the outcome measures?
We consider two alternative trial designs: a frequentist design, focussed on type-1 error control, and a Bayesian decision-theoretic design with a utility function to show that it controls the type-1 error rate and has suitable power properties to fit the regulatory environment. We show that in a frequentist design, we can switch to a group-sequential design with appropriate error control and perform an interim analysis, and we can consider changes to the trial operating characteristics without unblinding the data. We show that the Bayesian design can incorporate an unplanned interim analysis, and incorporate changes through appropriately specified loss functions.
The National Institute of Statistical Sciences Ingram Olkin Forum Series on Unplanned Clinical Trial Disruptions brought together groups of clinical trialists and statisticians to consider these challenges. This talk represents a collaborative position paper, and gives suggestions for further research in this area.

S19.4 The role of grace periods in comparative effectiveness studies of different medications

Aaron Sarvet[1], Jessica Young[2], Kerollos N. Wanis[3]

[1] *École Polytechnique Fédérale de Lausanne, Switzerland,* [2] *Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston MA, United States,* [3] *Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston MA, United States*

Many randomized and observational studies aim to compare the safety and/or efficacy of different pharmacological treatments indicated for the same disease diagnosis. In such studies, investigators might express interest in causal effects of sustained use of treatment. However, it is often unrealistic to expect study participants to adhere to a strict treatment strategy that demands continuous utilization. As an alternative, pragmatic randomized trials in practice often allow for \ textit{grace periods} - periods of time in which individuals take treatment as they would `naturally' until the end of the grace period, at which time they must take treatment if they have not already done so. However, estimates of these effects will not be transportable to settings in which natural treatment utilization patterns differ from those observed in the study, and interpreting the comparative effectiveness of two treatments will be challenging when their natural treatment utilizations differ.
We clarify the interpretation of effects under treatment strategies which include grace periods. As a compromise between strict protocols requiring continuous treatment utilization and natural grace period protocols, we introduce textit{stochastic} grace period strategies that mitigate some of the issues inherent in both aforementioned extremes. We consider assumptions needed for identification of this class of rules and how to reason about them under subject matter knowledge using Single World Intervention Graphs (SWIGs). We present worked examples using both simulated data and electronic health records data to illustrate how to implement statistical analysis for more transparent comparative effectiveness research allowing grace periods.

Parallel Sessions

S19.5 Analysis of trials with intervention induced post randomisation clustering

Dawn Teare[1], Stephen Walters[2], Matthew Bursnall[2], Jane Candlish[3]

[1] Newcastle University, United Kingdom, [2] ScHARR, University of Sheffield, United Kingdom, [3] BresMed

Many health interventions are delivered in a manner that may induce correlated outcomes for participants. This may occur via therapist effects or when participants are allocated to therapy groups in a non-random process. These designs are sometimes referred to a partially nested randomised controlled trials. This can present difficulty in the evaluation of such interventions in randomised trials as the correlation due to shared treatment factors may be confounded with correlation between participants pre-treatment due to the post-randomisation allocation process. The difficulty is most acute when evaluating the impact of proportionate interventions when for example a group-based intervention is offered only to the most severely affected participants.
Accounting for post-randomisation treatment induced clustering is not straightforward. Through two recent trial designs (with continuous outcomes) we demonstrate that accounting for such post randomisation clustering using mixed models may give biased estimates of the true treatment effects. We explore a number of methods to approach the analysis of such trials and propose some recommendations for special cases.

PARALLEL SESSION 20: Heterogeneity in effects

S20.1 Evaluation of mental health patients' diagnostic-therapeutic paths through state sequences analysis

Laura Savaré[1], Francesca Ieva[2]

[1] Politecnico di Milano, Italy, [2] MOX Department of Mathematics, Politecnico di Milano, Italy

BACKGROUND: In mental health, problems are frequently encountered in the quality of care provided. The care pathways of patients with severe mental disorders often do not correspond to the standards required by the evidence in this field. They present a high variability between countries and within them. This is mainly explained by the lack of definition of optimal diagnostic-therapeutic paths. In a real-world setting, the current tools resulted to be insufficient inadequately considering a phenomenon with such high variability.
OBJECTIVES: This work aims to develop innovative methods, based on the analysis of sequences, to i) monitor over time, ii) assess the quality, and iii) optimize the diagnostic-therapeutic pathways of patients with mental disorders.
METHODS: We perform sequence analysis of administrative data to identify different sequences of treatment followed by National Health Service beneficiaries resident in the Lombardy region with a diagnosis of schizophrenia. This methodology considers the patient's therapeutic path as a conceptual unit, i.e., a sequence, composed of a succession of different states that can describe patients in all their longitudinal aspects. In this work, we define the states to be the weekly coverage of different treatments (e.g., psychiatric visits, psychosocial interventions, and anti-psychotic drugs), and we use the Longest Common Subsequences (LCS) (dis)similarity measure to compare and group the sequences.
RESULTS & CONCLUSIONS: This technique is a valid support for describing complex scenarios, such as the heterogeneity of the paths followed by patients and possibly assessing the most efficient in preventing adverse events. We found common patterns of care that allowed us to risk profile patients. This kind of information can provide health policymakers an opportunity to plan optimum and individualized patient care by allocating appropriate resources, analyzing trends in the health status of a population, and finding the risk factors that can be leveraged to prevent the decline of mental health status at the population level. Therefore, the analysis of sequences is a precious tool for informing decision-making and supporting the definitions of new guidelines in a real-world evidence perspective.

S20.2 Causal DART: A non-parametric Bayesian approach to estimate heterogeneous treatment effects.

Ashwini Venkatasubramaniam[1], Karla DiazOrdaz[2], Chris Holmes[1,3],

[1] The Alan Turing Institute, United Kingdom, [2] London School of Hygiene and Tropical Medicine, United Kingdom, [3] University of Oxford, United Kingdom

Many medical and epidemiological studies involve studying the effects of a treatment or exposure of interest Z on an outcome Y. Increasingly, there is an interest to look beyond average effects, to understand heterogeneity of treatment with respect to observed characteristics. This has resulted in many methodological developments in the estimation of the so-called conditional average treatment effect (CATE), under the usual identification assumptions, $\tau(x) = E[Y \mid Z = 1,x] - E[Y \mid Z = 0,x]$. A very popular method, which has been shown to perform well in data-challenge competitions, is the Bayesian additive regression tree (BART). This data-adaptive method has been demonstrated to generate accurate estimates, detect complex interactions and require limited parameter tuning. It does so by modelling $E[Y \mid Z = z, x] = f(x, z)$ directly. This has the disadvantage that the prior distribution on $\tau(x)$ is induced only indirectly and is not able to adequately account for the uncertainty associated with future observations. Motivated by this, and inspired by the Bayesian bootstrap, we propose a Dirichlet prior BART, referred as causal DART, that generates a distribution over the BART estimated effects to better capture uncertainty. Both Causal BART and DART are sensitive to poor overlap, which can result in biased inferences, and we present extensions ps-BART and ps-DART by including the propensity score as a covariate in the models. A simulation study was conducted to compare the performance of these methods in terms of mean squared error, bias, 95% credible intervals coverage and interval length. The scenarios considered different sample sizes and good and poor overlap. We also considered the Bayesian Causal Forest (BCF) as a comparator. Our results show that Causal DART can outperform BCF (e.g., in the good overlap case, at a sample size of 750, coverage was 98% vs 90%). We also found that the BCF has similar coverage at n = 100 to Causal DART. BART and ps-BART have poor coverage at low sample sizes, but performance improves at n = 750. We illustrate the methods using a real-world dataset. Code for implementation is available as a R package.

S20.3 Adjusting for time of positive test when estimating the risk of a post-infection outcome

Shaun Seaman[1], Tommy Nyberg[1], Christopher Overton[2], David Pascall[1], Anne Presanis[1], Daniela De Angelis[1]

[1] University of Cambridge, United Kingdom, [2] Manchester University, United Kingdom

When comparing the risk of a post-infection binary outcome, e.g. hospitalisation, for two variants of an infectious pathogen (e.g. the SARS-Cov-2 virus), it is important to adjust for calendar time of infection. Typically, the infection time is unknown and the positive test time used as a proxy for it. Positive test time may also be used when assessing how the risk of the post-infection outcome changes over calendar time.
We show that if individuals' time from infection to positive test is correlated with their post-infection outcome, then the risk conditional on positive test time is a function of the trajectory of incidence of infection in the population. This means that the risk ratio adjusted for the positive test time can be quite different from the risk ratio adjusted for the infection time. We propose a simple sensitivity analysis method that indicates how the risk ratios adjusted for positive test time and infection time may differ. This method involves adjusting for a shifted positive test time, shifted to make the difference between it and the infection time uncorrelated with the post-infection outcome.
We illustrate this method by reanalysing data from a published study on the relative risk of hospitalisation following infection with the Alpha variant versus pre-existing variants of SARS-CoV-2. That study, which adjusted for positive test time, found that the risk was elevated for the Alpha variant. Our reanalysis indicates that the relative risk adjusted for infection time may be lower than the relative risk adjusted for positive test time.

S20.4 Disentangling interactions between components of complex health interventions

Silvia Metelli[1], Anna Chaimani[1]

[1] *Inserm Research Center of Epidemiology and Statistics, Université de Paris, France*

Networks of interventions can be used to summarise evidence from studies involving multiple treatments pertaining the same health condition, and network meta-analysis - an extension of standard meta-analysis that enables the simultaneous synthesis of all interventions in a single model - can then be employed to compare their effectiveness. Complex interventions are increasingly encountered in networks of studies. These interventions consist of multiple, potentially common and interactive components. When pooling results from studies involving complex interventions, the main interest does not only revolve around whether the intervention works per se, but rather what components contribute the most to the overall effectiveness and how they interact with each other. However, this is a statistically challenging problem as the effects of interacting components are hard to disentangle and may have complicated causal pathways. Furthermore, complex interventions tend to be very heterogeneous, due to increased variation across patient populations and interventions, which makes a reliable quantitative synthesis more challenging. In this work, we propose a model that considers each intervention component as a potential mediator in the path from treatment to outcome. We frame the mediating pathway as a novel Bayesian latent class model which allows to decompose each component effect into additive effects of distinct mediating paths when there exist multiple mediators (i.e. components here) that are correlated and interacting. A key underlying assumption of the model is to implicitly treat each combination of components as forming 'their own' class of interventions, and so the model may naturally incorporate interactions as class-specific parameters and better explain between-study variability. We illustrate our method using a network of 17 psychological interventions for coronary heart disease and we use synthetic data where specific pathways are hypothesised to show the contribution of each component on the total outcome effect. Extensions to Bayesian nonparametric latent class models can also be incorporated to investigate the benefits of allowing for more flexibility in the data generating mechanism of the class-specific parameters.

S20.5 Predicting individualized treatment effects using baseline risk: A simulation study

Alexandros Rekkas[1], Peter Rijnbeek[1], David M Kent[2], Ewout W Steyerberg[3], David van Klaveren[1]

[1] *Erasmus Medical Center, Rotterdam, Netherlands,* [2] *Tufts Medical Center, Boston MA, United States,* [3] *Leiden University Medical Center, Netherlands*

Our objective was to compare different risk-based methods for optimal prediction of individualized treatment effects from RCTs.
We simulated RCT data using diverse assumptions for the average treatment effect, a baseline prognostic index of risk (PI), the shape of its interaction with treatment (none, linear, quadratic or non-monotonic) and the magnitude of treatment-related harms (none or constant independent of the PI). The combination of these settings resulted in the definition of 648 simulation scenarios. In each sample we predicted absolute benefit using: models with a constant relative treatment effect; stratification in quarters of the PI; models including a linear interaction of treatment with the PI; models including an interaction of treatment with a restricted cubic spline (RCS) transformation of the PI with 3, 4, or 5 knots; an adaptive approach using Akaike's Information Criterion. We evaluated predictive performance using root mean squared error and measures of discrimination and calibration for benefit.
The linear-interaction model and the RCS-interaction (3 knots) model outperformed the constant treatment effect model in many simulation scenarios. The RCS-model was optimal when quadratic or non-monotonic deviations from a constant treatment effect were stronger, and when sample size was larger. Larger sample size also supported the use of the adaptive approach. Increasing the number of knots in the RCS models did not provide any improvement in predictive performance, while often tended to be overfitted, especially in scenarios with smaller sample sizes. All the simulation results can be explored at https://arekkas.shinyapps.io/simulation_viewer/. We also illustrated the application of the considered methods in the GUSTO-I trial which compared tissue plasminogen activator treatment to streptokinase in patients with acute myocardial infarction. The constant treatment effect method, the linear interaction method and the RCS with 3 knots had very comparable performance.
In conclusion, an interaction between baseline risk and treatment assignment generally improved treatment effect predictions. Non-linear interactions should be considered only in larger sample sizes.

PARALLEL SESSION 21: Communicating statistical concepts

S21.1 Why should I? Toward improved communication and evaluation of estimated dynamic treatment strategies

David Whitney[1], Stijn Vansteelandt[2], Karla Diaz Ordaz[1]

[1] *London School of Hygiene and Tropical Medicine, United Kingdom,* [2] *Ghent University, Belgium*

Identifying optimal treatment strategies for patients managing chronic health conditions is an area of importance in personalized medicine. The rapidly expanding literature on optimal dynamic treatment rules has historically emphasised estimating these strategies over communicating the benefit an individual patient can expect from following their optimal strategy. To this end we discuss both nonparametric and model-based effect measure estimands that, to varying degrees, summarize the expected benefit to an individual following the estimated care strategy. These estimands include (amongst others) the difference in the population-mean outcome under the estimated treatment rule as well as parametric summaries of conditional average treatment effects. For each estimand, we detail the construction of valid hypothesis tests and confidence intervals. Closely related to these approaches, we propose a non-inferiority test for comparison of simple, interpretable treatment rules to the unconstrained optimal treatment rule(s). We evaluate the proposed hypothesis tests and confidence intervals in simulations with both binary and multi-category treatments. We apply the methods to identify (approximately) optimal treatment intensification strategies for type 2 diabetes patients already taking metformin using data from the Clinical Practice Research Datalink, which consists of electronic health records for primary care patients in the United Kingdom.

S21.2 P-value, s-value, b-value, d-value,... What else? Individual Success Probability

Bernard Francq[1], Ron Kenett[2]

[1] *GSK, Belgium,* [2] *KPA*

In medical research, statistical significance is often based on confidence intervals (CIs) and p-values, the reporting of which is included in publications in most top-level medical journals. However, recent years have seen ongoing debates on the usefulness of these parameters, leading to a significance crisis. Misinterpretations of CIs and p-values can lead to misleading conclusions and nonreproducable claims. The lower the p-value does not necessarily mean the better the treatment.
The s-value or B-value (namely the probability that a patient under treatment A ends up with a better clinical outcome compared to an another patient under treatment B) have been proposed in the literature as alternative solutions.
Here, we promote the concept of individual success probability (ISP), which is a wider and generalized definition based on the concept of tolerance intervals. The ISP allows a clear interpretation following both frequentist and Bayesian paradigms. Using synthetic examples with the 1-sample or 2-samples t-test, we show that the lower (or upper) bound of the ISP is a one-to-one function of the p-value with enhanced interpretability properties. The lower (upper) bound of the ISP has a default cut-off value of 50% whatever the type I error that avoids the common pitfalls of the CIs and p-values. The ISP offers enhanced insights in reviewing statistical analysis in medical research from such a perspective. An orthopedic surgery study is used to illustrate the ISP on a mixed model. We argue that the ISP should be preferred in clinical trials by researchers and considered by journal editors.

S21.3 Statistical advising: professional development opportunities for the biostatistician

Marissa LeBlanc[1], Corina S Rueegg[2], Nural Bekiroğlu[3], Tonya M Esterhuizen[4], Morten W Fagerland[2], Ragnhild S Falk[2], Kathrine Frey Frøslie[5], Erika Graf[6], Georg Heinze[7], Ulrike Held[8], René Holst[2], Theis Lange[9], Madhu Mazumdar[10], Ida Hed Myrberg[11], Martin Posch[7], Jamie C Sergeant[12], Werner Vach[13], Eric A Vance[14], Harald Weedon-Fekjær[2], Manuela Zucknick[1]

[1] University of Oslo, Norway, [2] Oslo University Hospital, [3] Marmara University, Istanbul, Turkey, [4] Stellenbosch University, South Africa, [5] Norwegian University of Life Sciences, As, Norway, [6] University of Freiburg, Germany, [7] Medical University of Vienna, Austria, [8] University of Zurich, Switzerland, [9] University of Copenhagen, Denmark, [10] Icahn School of Medicine at Mount Sinai, New York NY, United States, [11] Karolinska Institutet, Solna, Sweden, [12] University of Manchester, United Kingdom, [13] Basel Academy for Quality and Research in Medicine, Switzerland, [14] University of Colorado, Boulder CO, United States

In contrast to the theoretical statistician, the biostatistician needs broad knowledge and understanding of applied statistics, which has led to the development of specific training programs for medical statistics and biostatistics over the last 50 years. In addition, the job of the biostatistician is multi-disciplinary and collaborative in its essence, meaning that training and practical experience in interdisciplinary research and communication of statistics to biomedical researchers is required. We recently published a commentary (LeBlanc et al 2022) where we argue that biostatistical advising should be considered a key activity in the biostatistician's professional development and with the appropriate frameworks is beneficial for both the biostatistician and the biomedical research process.
We describe the three types of advising 1) advising in a single or a few advising session(s), 2) single-project collaboration, and 3) long-term collaboration, which all have their different opportunities for the biostatistician's development. We think that the academic institution organising the advising has the potential to both promote career development and at the same time increase the efficiency and quality of the statistical advice by good prioritisation and matching of advising projects.
One of the most important requirements of successful interdisciplinary collaboration, in our opinion, is mutual respect. All collaborators should recognize each other as important contributors to the scientific process and success of a project. Organisers of statistical advising can help to achieve this by framing their centre as a place of contact for initiating new research collaborations and highlighting the research activities of the biostatisticians in the centre.
We conclude that, equally to teaching and research, biostatistical advising should be seen as an essential part of the foundation on which to build a biostatistician's professional development. We see advising as one of the main arenas for developing collaborative research skills, establishing new collaborative research projects and promoting problem-driven methods development.
This abstract is based on and uses excerpts from our newly published commentary: LeBlanc, M, et al. «Statistical advising: Professional development opportunities for the biostatistician.» Statistics in medicine 41.5 (2022): 847-859.

S21.4 The inflation of p-values of likelihood-ratio tests in longitudinal data analysis

Dominik Grathwohl[1]

[1] Société des Produits Nestlé, Switzerland

INTRODUCTION: One of the most frequently asked questions about lme4 is «how do I calculate p-values for estimated parameters?» The R Documentation continues: Users who need p-values have a variety of options, for likelihood-ratio tests via anova. When fitting splines with mixed models, this option badly fails. In this simulation study, we investigated different longitudinal models along a dataset, where repeated measurements were crossed over. We investigated 3 longitudinal models using splines and one mixed model where time was category for comparison purposes.
METHODS: A dataset was simulated mimicking published microbiota data (David et al.). The dataset had 11 test days which were crossed over by two periods. The dataset had variance heterogeneity within period, but homogeneity between periods. The correlation within period was 0.5 and between periods 0.2. The mean profile was mimicking the published data of David et al. Missing values were introduced as observed in David at al.. For estimating from the null-hypothesis, the treatment was permutated. The 3 longitudinal models were: 1. gamm with thin plate spines and a random intercept; library(mgcv), 2. generalized least squares (gls) with restricted cubic splines and spherical correlation structure; library(nlme) and library(rms), 3. lmer with restricted cubic splines (lmer-rcs) and random intercept; library(lme4) and library(rms). The mixed model where time was category was realized with lmer and had a random intercept. All models were fitted in maximum likelihood mode. Inflation of the p-value is present if the frequency of rejecting the null is greater than 0.05 under the null-hypothesis.
RESULTS: The inflation of the p-values were 0.96, 0.17, 0.18, 0.15 for gamm, gls, lmer-rcs, lmer, respectively.
DISCUSSION: Investigating time trends in cross over settings is a quite common task. Apparently, the likelihood-ratio under the null-hypothesis does not follow a chi-square distribution with the respective degrees of freedom. We recommend as a remedy bootstrapping or permutating the dataset to obtain an empirical null-distribution of the likelihood-ratios for statistical inference.
REFERENCES: David et al. "Diet rapidly and reproducibly alters the human gut microbiome", Nature 2014

S21.5 Visualising Master Protocols

Deepak Parashar[1]

[1] University of Warwick, Coventry, United Kingdom

One of the challenges facing personalised cancer therapy is to effectively and appropriately target multiple treatments for sensitive patient population based on a panel of biomarkers across multiple cancer types, and matched clinical trial designs. Of late, there has been a surge in 'Master Protocols' as novel biomarker-guided trial designs comprising multiple tumour types, multiple mutations and multiple drugs, a methodology that best seems to address this challenge. Whilst there have been great strides in its statistical development and clinical implementation, some key issues around the structure of master protocols are often either overlooked or are somewhat less well-understood.
Biomedical data, especially genomic data highlighting prognostic and predictive potential of biomarkers, are inherently high-dimensional. Therefore, due to the large number of tumour types and mutations, data collected at the end of a clinical study based on master protocol design will no longer be small but potentially large and multidimensional and hence standard statistical techniques could be of limited use. In order to gain insight and draw hypothesis-driven inferences from such high-dimensional genetic datasets, one risks missing out on the shape and structure of the overall data. While 'small' datasets can be analysed using summary statistics to get an impression of the data, summarising high-dimensional data using tables and descriptive statistics is not adequate. Visualisation has the potential to 'bridge the gap' between complex data and the human who needs to understand the data. Hence, appropriate visualisation techniques are required for analysis of multidimensional biomedical data such as master protocols.
We present novel geometric methods (mathematical theory of hypersurfaces) to visualise complex cancer therapeutic data as multidimensional. In particular, master protocol designs are cast as hypersurfaces by assigning a dimension to each variable, namely, the treatments, biomarkers, and disease classifiers. Basket and Umbrella trial designs then emerge as embedded Euclidean subspaces via projection of such hypersurfaces. The framework has the potential to offer oncologists visual guidance in appropriately targeting multiple therapies for sensitive groups of patients with flexibility to incorporate further evidence from clinico-genomic or multi-omic data.

PARALLEL SESSION 22: Counterfactuals

S22.1 Counterfactual simulation to evaluate sequential stratification methods

Ilaria Prosepe[1], Hein Putter[1], Nan van Geloven[1]

[1] Department of Biomedical Data Sciences, Leiden University Medical Centre, Netherlands

Estimating counterfactual survival outcomes with and without treatment from observational data poses many challenges. This is especially true when treatments are initiated over longitudinal follow-up. In that case, causal adjustments methods used to correct for confounding need to respect the longitudinal structure of the problem at hand. In particular, the time axis relevant for creating comparability between treatment groups (e.g. time since study entry) may differ from the time axis at which time treatment decisions are made (e.g. calendar time). An approach that allows accounting for these two time axes is sequential stratification, i.e. comparing treatment initiators and non-initiators across a sequence of artificially constructed time origins, restricting to individuals at risk.
In this work, we perform an extensive simulation study assessing the accuracy of the sequential stratification method, with tailored causal adjustments, for estimating survival benefit for patients on the waiting-list for liver transplantation. In this setting, decisions on treatment initiation are made in calendar time, when new livers become available, and survival benefit needs to be estimated at specific calendar dates on cross-sections of patients rather than a cohort of patients. Moreover, some important time-dependent predictors of death are also highly predictive of treatment initiation. Our aim is to estimate counterfactual survival outcomes, defined as restricted mean survival times (RMST), and their contrast at any given calendar time for all patients on a waiting list for treatment. In alignment with this aim, we wish to assess the performance of different modeling strategies at a set of evenly spaced calendar dates (cross-sections). This implies generating the counterfactual RMSTs at each cross-section in a consistent way for each simulated patient. In this simulation of counterfactuals, we generate complete synthetic longitudinal data comprising of time-to-death while waiting for treatment, time-to-treatment, post-treatment survival. From this data we derive the "true" counterfactual RMSTs. We then evaluate the accuracy of the sequential stratification method by comparing the "true" RMSTs with the estimates provided by the method, at each cross-section for all patients that would have been alive and waiting for treatment, had treatment never been administered.

Parallel Sessions

S22.2 Bias in multivariable Mendelian randomization studies due to measurement error on exposures

Andrew Grant[1], Jiazheng Zhu[1], Stephen Burgess[1]

[1] *University of Cambridge, United Kingdom*

Multivariable Mendelian randomization estimates the causal effect of multiple exposures on an outcome, typically using summary statistics of genetic variant associations taken from separate samples. Exposures of interest in Mendelian randomization applications will often be measured with error. The summary statistics will therefore not be of the genetic associations with the exposure, but with the exposure measured with error. Classical measurement error will not bias genetic association estimates but will increase their standard errors. With a single exposure, this will result in bias toward the null in a two-sample framework. Thus, two-sample Mendelian randomization still provides a valid test of the causal null hypothesis regardless of measurement error. However, this will not necessarily be the case with multiple correlated exposures.
We shall demonstrate how the direction and size of bias, as well as coverage, power and type I error rates in multivariable Mendelian randomization studies are affected by measurement error on exposures. We present a method to account for measurement error using a maximum likelihood framework which can be performed using a fast and easily implemented algorithm. Finally, we shall present the results of an applied example which suggests that measurement error leads to the effect of body mass index on coronary heart disease risk to be overestimated, and that of waist-to-hip ratio to be underestimated, using a conventional multivariable Mendelian randomization analysis.

S22.3 Sequential counterfactual prediction to support individualized decisions on treatment initiation

Pawel Morzywolek[1], Johan Steen[2], Wim Van Biesen[2], Johan Decruyenaere[2], Stijn Vansteelandt[1]

[1] *Ghent University, Belgium,* [2] *Ghent University Hospital, Belgium*

Digitalization of patient records and increasing computational power have led to a paradigm shift in the field of medical decision-making from one-size-fits-all interventions to data-driven intervention strategies optimised for particular sub-populations or individuals. In this research, we make use of statistical techniques on counterfactual prediction and heterogeneous treatment effects (CATE) estimation to support patient-centred treatment decisions on the initiation of renal replacement therapy.
In particular, we applied sequential counterfactual prediction strategies to help physicians in making patient-centered decisions on treatment initiation. For this, we made use of linked databases from the Intensive Care Unit (ICU) and dialysis center of the Ghent University Hospital, containing longitudinal, highly granular records from all adult patients admitted to the ICU from 2013 to 2017. Based on these data, we inferred whether or not to initiate renal replacement therapy (RRT) for individual patients suffering acute kidney injury (AKI) on each day of their ICU stay, based on their measurements until that day.
Because RRT is not a feasible treatment option for many patients, standard counterfactual prediction strategies are fallible. In our analysis we therefore study the use of retargeted causal learning techniques, which reweight individuals to generate a population of patients for whom all treatment options are plausible, so that counterfactual prediction becomes feasible. The proposed approach leverages recent developments on orthogonal statistical learning literature, resulting in a class of Neyman-orthogonal weighted loss functions for counterfactual prediction. Our approach is guaranteed to deliver predictions in the support of the counterfactual outcome mean, and delivers oracle behaviour due to orthogonality of the loss function (where orthogonality is relative to the infinite-dimensional propensity score and conditional outcome mean).
Our study involves a performance evaluation of state-of-the-art methods from the literature on heterogeneous treatment effects estimation, e.g. DR-Learner and R-Learner, which can be viewed as special cases resulting from the proposed retargeting framework for counterfactual prediction and CATE estimation.

S22.4 Assessing discrimination of counterfactual prediction models for time-to-event outcomes

Nan van Geloven[1], Ruth Keogh[2]

[1] *Department of Biomedical Data Sciences, Leiden University Medical Centre, Netherlands,* [2] *The London School of Hygiene & Tropical Medicine, United Kingdom*

Counterfactual prediction models provide estimates of absolute risks under a particular treatment pattern while conditioning on other patient characteristics that are predictive of the outcome. Counterfactual predictions model can inform treatment decisions by providing estimates of a patient's risk if they were to be given the treatment and their risk if they were not to be given the treatment. This is an important improvement over 'factual' prediction models that typically ignore treatments during model development and then only generate risks that apply to the mix of treated and untreated patients as observed in the development dataset.
While the importance of counterfactual predictions is recognized, concerns have been raised that counterfactual prediction models cannot be validated in 'factual', i.e. observational, datasets. Thorough validation is pivotal for any prediction model before it is used in medical practice, including assessment of the discriminative ability of the model. For time-to-event outcomes that means assessing whether the model assigns higher risk estimates to patients who will experience the event earlier than others. In this work we propose extended versions of two popular measures of discrimination for time-to-event outcomes, the c-index and the cumulative/dynamic area under the receiving operator characteristic curve (C/D AUCt), that allow validation of counterfactual prediction models. Our focus is on time-to-event outcomes and on a binary time-dependent treatment. We assume a validation dataset exists in which patients are observed with different treatment patterns over time, and allow for time-dependent confounders. Our aim is to assess the discriminative value of counterfactual risk predictions under a particular fixed treatment pattern. We artificially censor patients when they first deviate from this treatment pattern and construct time-dependent inverse probability of censoring weighted versions of c-index and cumulative/dynamic AUCt. Validity of the proposed performance measures is examined through counterfactual simulations. In these we mimic different scenarios, varying the strength of the treatment effect and predictors, extent of time-dependent confounding and observed treatment patterns. Our results show that validation of counterfactual predictions is indeed possible with the extended discrimination measures.

S22.5 Estimation and calibration of counterfactual risk predictions, with application to liver transplant

Ruth Keogh[1], Ilaria Prosepe[2], Nan van Geloven[2]

[1] *The London School of Hygiene & Tropical Medicine, United Kingdom,* [2] *Department of Biomedical Data Sciences, Leiden University Medical Centre, Netherlands*

Clinical risk prediction models enable predictions of a person's risk of an outcome (e.g. mortality) given their observed characteristics. It is often of interest to use risk predictions to inform whether a person should initiate a particular treatment. However, when standard clinical prediction models are developed in a population in which patients follow a mix of treatment strategies, they are unsuitable for informing treatment decisions. Counterfactual risk predictions (CRPs) aim to address this problem. CRPs are estimates of what a person's risk would be if they were to follow a particular treatment strategy, given their individual characteristics that are also predictive of the outcome. CRPs do not depend on the observed mix of treatments and are therefore suitable for informing treatment decisions.
This work describes methods for estimating CRP models using longitudinal observational data on treatment use, patient characteristics and a time-to-event outcome. We use causal inference methods based on marginal structural models and censoring-and-weighting to handle treatment switching and time-dependent confounding. In our motivating example, interest lies in CRPs for mortality risk in patients awaiting a liver transplant under the strategies of receiving or not receiving a transplant.
A crucial step in validating prediction models is assessing calibration, which involves comparing a person's predicted risk with their observed outcome. However, calibration methods for standard prediction models do not extend directly to CRPs. This is because CRP involves obtaining predictions under a treatment strategy that may differ from that observed for a given patient, making comparison of predicted and observed outcomes difficult. We will present new methods for assessing calibration of CRPs. One approach uses inverse probability weighting to estimate 'observed' risks under the treatment strategy of interest and compares these with the CRPs. An alternative approach uses the CRP model to obtain a risk prediction for each patient under their observed treatment pattern, and compares this with their observed outcome at the end of follow-up. We will illustrate these approaches using simulation studies and present results from our motivating example in which we use US Scientific Registry of Transplant Patients data to fit a CRP model and assess its calibration.

Parallel Sessions

PARALLEL SESSION 23: Bias and Estimation

S23.1 Estimation of treatment effects in randomized clinical trials involving external control data

Heiko Goette[1], Marietta Kirchner[2], Johannes Krisam[2], Arthur Allignol[3], Francois-Xavier Lamy[4], Armin Schueler[1], Meinhard Kieser[2]

[1] Merck Healthcare KGaA, Darmstadt, Germany, [2] Institute of Medical Biometry, University of Heidelberg, Germany, [3] Daiichi Sankyo Europe GmbH, [4] UCB Pharma

RCTs are the gold standard for assessing new treatments as they provide unbiased treatment effect estimates. However, in early clinical development low sample sizes lead to high variability in estimates from RCT. The variability could be reduced by leveraging external control data (augmented RCT) which might introduce bias in return. For the common setting of subject-level control group data available from one previous clinical trial or real-world data (RWD) source, we evaluate different analysis options for augmented RCTs.
Internal RCT and external control data can be combined via outcome or baseline covariate data. We focus on Bayesian hierarchical models (BHM) for the outcome and propensity-score (PS) based approaches for baseline covariates. The former covers observable and unobservable confounders in a non-specific way while the latter focuses explicitly on observable confounders. A combination of both is desirable: substantial differences in covariate distributions raises doubts about similarity of underlying populations and appropriateness of borrowing outcome data; however, PS-based methods cannot identify substantial differences in outcomes caused by unobservable confounders. One option is covariate adjusted BHM. We propose a new option: a model averaging-based combination of BHM estimates and PS-based estimates is used if the covariate data from in- and external data sources are comparable as assessed by a PS-based criterion. Otherwise, external control data is ignored and the internal RCT estimate is used.
In a simulation study the approaches are compared in terms of bias and RMSE for the hazard ratio on progression-free survival-type data. Varying assumptions regarding (un-)observable confounder distributions and assessment schedules are investigated. This allows to cover scenarios which resemble either external controls from previous clinical trials or RWD.
Our simulation study suggests that for external clinical trial data, the proposed approach is a viable option which performs in most scenarios better than the other investigated options. With increasing differences in terms of unobservable confounders and assessment of disease progression between internal RCT and RWD, the advantage of the augmented RCT analysis options decreases compared to the internal RCT estimate.

S23.2 Performance of different estimators in adaptive two-stage trials with optimized design parameters

Jan Meis[1], Maximilian Pilz[1], Meinhard Kieser[1],

[1] Institute of Medical Biometry, University of Heidelberg, Germany

The flexibility of adaptive clinical trial designs can offer significant advantages over fixed designs. If design parameters are chosen appropriately, adaptive designs can decrease the expected sample size and save resources in cases where there is early evidence that the continuation of a trial may be futile. A known issue with adaptive designs is that estimators appropriate in a single-stage fixed design setting, such as the maximum likelihood estimator, can be biased due to the dependence structure in the data introduced by the adaptivity. This problem affects point estimators as well as confidence intervals and p-values. Regulatory agencies such as the EMA or FDA recognize this problem and urge researchers that «the extent of bias should be evaluated, and estimates should be presented with appropriate cautions regarding their interpretation» [1].
Over the years, various methods have been proposed to mitigate the bias introduced by adaptive designs. However, estimators often need to fulfill other requirements to be useful in practice, such as having an acceptable variance. In this work, we provide results on the operating characteristics of different estimators in optimal adaptive two-stage designs with normally distributed outcome. In optimal designs, design parameters such as the sample sizes, decision boundaries, and adaptation rules are chosen as the result of an optimization process. The goal of the optimization process is to maximize some metric of design quality, a typical example being the expected sample size required to fulfill certain power requirements under a specified hypothesis. Although optimal adaptive designs have been a topic of recent research, guidance on estimation in this novel kind of designs is still scarce. We compare classical and recently developed point estimators as well as estimation methods for confidence intervals and p-values regarding various performance criteria such as bias, variance, mean squared error, coverage, and consistency with test decisions.

REFERENCE: [1] FDA (2019). Adaptive Designs for Clinical Trials of Drugs and Biologics. Guidance for Industry. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adaptive-design-clinical-trials-drugs-and-biologics-guidance-industry

S23.3 Considerations for the design and analysis of nested case-control studies

Monisha Dewan[1], Holly Tovey[1], Lucy Kilburn[1], Nuria Porta[1], Maggie Chon U Cheang[1], Emma Hall[1], Judith Bliss[1]

[1] Clinical Trials and Statistics Unit, The Institute of Cancer Research, London, United Kingdom

MOTIVATION: Given increased interest in personalized medicine, efficient methods are required to identify and evaluate biomarkers predictive of outcomes and treatment effects. Many oncology trials embed sample collections within the trial protocol and these tissue collections provide a useful resource for biomarker research. However, despite advancements in high-throughput technologies, omics analysis of a large cohort of samples may be infeasible due to prohibitive cost.
METHODS: Nested case-control studies involve the identification of cases (patients with the event/outcome of interest) and matched controls, allowing analysis in a smaller subset enriched for events of interest. When designed and conducted appropriately, these provide an alternative approach to analysis of the full cohort while maximising use of existing data and being more efficient than standard cohort designs, particularly in the presence of rare events. Efficiency and validity of these studies is impacted by choice of control sampling, matching strategy, and data analysis methods. These choices also determine event rates within the sub-study meaning caution is required when interpreting estimates. While risk set sampling (selecting controls from patients who are event free at the time of an event) can prevent selection bias as it is not reliant on future information, cumulative incident sampling (only considering those who are event free at the end of the study period) may be more appropriate when the likelihood of an event is very low ("rare disease assumption"). Sampling with replacement (selected controls are replaced back into the at-risk group for potential re-selection) has been shown to be unbiased and can further reduce the number of samples to be processed. However, this may not seem intuitive to collaborators and may require multiple pairs to be dropped/replaced if a single sample is non-evaluable. Case-control matching ratio will have an impact on required resources as well as resulting event rates within the sub-study, and selection of variables to match on will have implications for what effects can be tested for.
SUMMARY: We will discuss the main methodological considerations when running a nested case-control study from the perspective of an oncology based clinical trials unit with motivating examples from our own portfolio and simulations.

S23.4 Confidence Band for the Cumulative Hazard Rate Function in Right Censored Length-biased Sampling

Ali Shariati[1], Hassan Doosti[1], Vahid Fakoor[2],

[1] Macquarie University, Australia, [2] Ferdowsi University of Mashhad, Iran

To study the natural progress of a disease, a prevalent cohort study design is more practical and efficient than an incident cohort design as it could shorten the period of study considerably. In the former, cases who have experienced the initiating event (e.g., the onset of disease) prior to the commencement of the study are recruited. Then, the recruited subjects are typically followed-up for the remaining of their lifetimes until experiencing the terminating event (e.g., death for example). Such a sample does not comprise a random collection of the target population because of structural sampling bias. A very common case of this bias which occurs when the initiating events follow a stationary Poisson process is called length-bias. The survival data collected in prevalent cohort studies are commonly analysed by conditioning on observed truncation times. Under length-bias, however, we can employ a more efficient method known as unconditional approach. The other frequent challenge we encounter in prospective cohort studies is loss to follow-up on some of the subjects (for variety of reasons) which leads to informative censoring.
In the Canadian Study of Health and Aging (CSHA), a random sample of elderly individuals aged 65 and above throughout Canada was screened for cognitive impairment. For those subjects diagnosed with two types of dementia (vascular dementia and Alzheimer's disease), the date of onset was ascertained, and a five-year period of follow-up was carried out. The incidence rate has been reported by previous studies to be stationary, and thus the collected data are length-biased and right censored. We use the unconditional nonparametric maximum likelihood estimator (MLE) of the cumulative hazard rate function to estimate the cumulate risk at any specific age. We then review the asymptotic properties of the estimator proposed. After that, we derive a nonparametric statistic based on the unconditional MLE. The quantiles of the limiting distribution of the statistic are obtained, which are then employed to derive uniform confidence bands for the cumulative hazard rate function. The statistical tools are applied to analyse simulated data and the observations on patients with dementia.

S23.5 Selection bias and multiple inclusion criteria in observational studies

Stina Zetterstrom[1], Ingeborg Waernbaum[1]

[1] *Uppsala University, Sweden*

Spurious associations between an exposure and outcome not describing the causal estimand of interest can be the result of selection of the study population. Recently, sensitivity parameters and bounds have been proposed for selection bias, along the lines of sensitivity analysis previously proposed for bias due to unmeasured confounding. The basis for the bounds is that the researcher specifies values for sensitivity parameters describing associations under additional identifying assumptions. The sensitivity parameters describe aspects of the joint distribution of the outcome, the selection and a vector of unmeasured variables, for each treatment group respectively. In practice, selection of a study population is often made on the basis of several selection criteria, thereby affecting the proposed bounds. In this paper we extend the previously proposed bounds to give additional guidance for practitioners to construct i) the sensitivity parameters for multiple selection variables and ii) an alternative assumption free bound, producing only logically feasible values. Our results show that the assumption free bounds can be both smaller and larger than the previously proposed bounds and therefore can serve as an indicator of settings when the former bounds do not produce feasible values. As a motivating example we derive the bounds for causal estimands in a study of perinatal risk factors for childhood onset type 1 diabetes mellitus where selection of the study population was made by multiple inclusion criteria. It may be difficult for the researcher to give plausible input values for the sensitivity parameters for selection bias under multiple selection and to provide further guidance for practitioners, we provide a data learner in R where both the sensitivity parameters and the assumption free bounds are implemented.

PARALLEL SESSION 24: Lightning never strikes twice

S24.1 Surrogate endpoints in regulatory use: how many are actually statistically valid?

Wang Pok Lo[1], Oriana Ciani[2], Steff Lewis[1], Hannah Ensor[1], Christopher Weir[1]

[1] *Usher Institute, University of Edinburgh, United Kingdom,* [2] *SDA Bocconi School of Management, Mila, Italy*

Surrogate endpoints replace clinical endpoints, and are thus expected to predict the clinical benefit or risk of an intervention. Biomarkers or earlier measures of clinical benefit are often used as surrogates. With their increasing diffusion linked to accelerated regulatory pathways, it is vital that surrogate endpoints have been statistically and clinically validated. This involves showing that the treatment effect on a surrogate endpoint has a sufficient and reliable association to that on the clinical endpoint.
The majority of recent literature findings share one key result: most surrogate endpoints approved by regulators are statistically inadequate when assessed under existing validation frameworks. However, most of these findings have either been limited to one or two therapeutic areas such as oncology, or endpoints approved by one regulatory agency.
Therefore, a systematic review is being undertaken to explore i) the prevalence of validated surrogate endpoints across all therapeutic areas, and ii) the methods used for validation. Five electronic databases were searched for surrogate validation studies, yielding 18,904 hits. In terms of regulators, the websites of United States Food and Drug Administration (FDA) and European Medicines Agency (EMA) were searched for instances of surrogate endpoints supporting drug approvals.
Literature screening in both the title/abstract and full-text stages will be completed by two independent reviewers. Items to be extracted include names of surrogate and primary endpoints, statistical measures of surrogacy evaluation, whether the surrogate was valid and whether FDA or EMA has approved the surrogate endpoint of interest; consequently, non-validated surrogate endpoints in regulatory use will also be identified.
Data analysis will be presented through a tabular list and a narrative synthesis. The status of validated surrogate endpoints will be summarised by scores described in two surrogate validation frameworks (IQWiG and BSES3) which will be compared to the surrogate lists in FDA and EMA. Data will be grouped by framework score, therapeutic area and statistical method of evaluation to identify any patterns in the validation status.
This presentation will report on the full results to be obtained from the systematic review, which will act as a comprehensive list to inform statisticians and regulators.

S24.2 Application of the R Shiny app DetectoR for signal detection and label generation

Martin Gebel[1], Ann-Kathrin Frenz[1], Tobias Machewitz[1], Steffen Jeske[2], Antigoni Elefsinioti[1], Amrit Singh[3], Annette Boeckenhoff[1]

[1] *Bayer AG, Germany,* [2] *Chrestos Concept GmbH & Co. KG, Germany,* [3] *The University of British Columbia, Vancouver BC, Canada*

Safety analyses in clinical trials are continuously growing in importance. This triggers the requirement to put more sophistication and statistical consideration into their application. The availability of plentiful details in safety data calls for a clear overview of dense information and an efficient way to provide data insights.
The DetectoR R Shiny application provides a handy platform allowing for early identification of signals and an ongoing monitoring of safety along the medical product development lifecycle using CDISC standard SAS analysis datasets using state-of the art methodology.
The first highlight of DetectoR is the display of a double dot plot showing for each adverse event of interest the incidence proportion per treatment group and risk ratio as effect estimate. Interactively, incidence rates and risk difference can be chosen as event measure and effect estimate, respectively. The user can choose from different techniques for p-value adjustment of multiplicity, including calculations based on the False Discovery Rate (FDR) and the new Double FDR with flexible alpha levels. Additionally, DetectoR not only supports the analysis of single studies, but also integrated analyses of studies with the option of supplying study stratified estimates.
Moreover, generic data filtering is possible based on subject level characteristics and all available adverse event categories.
A heatmap based on the MedDRA hierarchy presents the second highlight of DetectoR. It provides an appealing and interactive overview of the adverse events' distribution across different groupings. With this graphical display the users can navigate through the MedDRA hierarchy in various ways by zooming in and out of the SOC and subcategories. Generic filtering and a color coding either based on p-values or effect estimates, allows the users to adjust the heat map according to their particular needs.
The application of DetectoR in previous submissions at Bayer has supported the label creation and simplified the identification of label-relevant adverse events.

S24.3 Internal pilot designs for sample size recalculations in animal experiments - prospects and pitfalls

Michael Lauseker[1]

[1] *Ludwig-Maximilians-Universität München, Germany*

When planning an animal experiment, the statistician is often confronted with the challenge that the investigator has a conception of a meaningful difference between two groups, however not even an idea on standard deviations is present. Due to logistical and legal aspects, a proper pilot study is often infeasible. Methods for sample size recalculation using an internal pilot study could offer a solution in this situation, as they operate in the spirit of the so-called 3Rs, especially "Reduction". These methods have been widely studied in clinical trials settings, however their performance and usability in animal experiments remain unclear. Besides regulatory aspects, the main differences is that effect sizes in animal experiments are larger and sample sizes are therefore usually considerably smaller than in clinical trials.
In this contribution, several methods are examined, restricted to continuous outcomes. Due to the unfavourable characteristics of blinded variance estimation when dealing with small samples and possibly large differences, this contribution focuses on unblinded sample size recalculations.
A few adaptions have to be made for animal experiments: A fixed small number of animals is used for the internal pilot study and a maximum number has to be defined. After analysing these first animals, the sample size recalculation is performed. Based on simulation studies for different effect sizes, sample sizes for the internal pilot study, the maximum sample size and the recalculation procedure, power and type I error rate are investigated and recommendations are given.
It could be shown that all methods have some drawbacks, especially when the effect sizes are large or the maximum number is rather low. But even in these situations, they could offer an improvement to the current situation where regularly no a priori power and sample size analysis is conducted. To maintain scientific rigour, it is however essential that the analysis procedure is accurately prespecified in an analysis plan.

S24.4 Pragmatic SIMEX method to correct for measurement error in time-varying prescription-based exposures

Steve Ferreira Guerra[1], Michal Abrahamowicz[1], Robert Platt[1]

[1] *McGill University, Monreal QC, Canada*

Drug prescription registries have been increasingly used in pharmaco-epidemiological drug safety and efficacy studies to determine individual drug exposures. However, drug exposures based on prescriptions may not accurately represent the actual drug intake. This may for example happen because of non-adherence to the prescribed regimen. Such discrepancies induce measurement error (ME) in the assessment of the true drug exposure, which is known to result in biased inference in naive analyses.
We propose to develop a method specifically tailored to correct for ME in the analyses of time-varying prescription-based exposures. The proposed approach relies on a pragmatic adaptation of the simulation-extrapolation (SIMEX) method. SIMEX is a well-known bias correcting method which requires some distributional assumptions about the ME. Since the mechanism from which the ME arises may be complex in this context, the proposed method additionally avoids any specification about the true ME, using instead an observable function of the data as a proxy for ME.
We will conduct a simulation study to evaluate to what extent the proposed method reduces bias caused by this type of ME. Observed prescription histories will be resampled from a real pharmaco-epidemiological dataset and true drug intake simulated under various plausible assumptions about non-adherence patterns. We will also investigate the behavior of the method with respect to different extrapolation functions, which might affect the performance. Preliminary results indicate that the proposed method performs well in simulation, by reducing the bias due to ME.
In conclusion, accounting for ME in time-varying prescription-based exposures is challenging, but crucial to avoid biased analyses. Pragmatic methods, such as the one proposed, are a promising avenue to more appropriately account for complex types of ME.

S24.5 Estimating the effects of multiple treatments in combination on outcomes using longitudinal data

Emily Granger[1], Ruth Keogh[1]

[1] *The London School of Hygiene & Tropical Medicine, United Kingdom*

Recent years have seen considerable development in generalised methods (g-methods) for estimating the causal effects of time-varying treatments or exposures on outcomes using longitudinal observational data when there is time-dependent confounding. Most research on g-methods has focused on settings with a single binary treatment of interest. There is a lack of research investigating the performance of g-methods to estimate effects of multi-level or combined exposures, e.g. representing multiple treatment combinations. We aim to compare three g-methods for use in this setting: inverse-probability-of-treatment weighted estimation of marginal structural models (IPTW), the g-computation formula and g-estimation of structural nested models. Methods are compared in a simulation study and a motivating applied example.
Our simulation study focused on estimating the combined effects of two binary time-varying treatments (T1, T2) on a time-varying continuous or binary outcome, all measured at 5 time points. There were four treatment strategies of interest defined by the four possible combinations of T1 and T2. Our focus was on sustained treatment strategies up to time horizons 1-5 years and the estimands of interest were mean differences in the outcome between treatment strategies. The simulation included scenarios with varying effect sizes, prevalence of T1 and T2, and amount of time-dependent confounding. We also included scenarios with and without interaction effects between the two treatments, and short- and long- term effects.
All g-methods gave unbiased estimates in most scenarios. The main exception was IPTW in the scenario with a strong association between treatment and confounder. For all methods, the variance increased as time increased, as expected. On average, IPTW was the least efficient method with particularly large variance in scenarios with low treatment prevalence or strong exposure-confounder associations.
Our motivating example comes from cystic fibrosis. Many people with cystic fibrosis are prescribed at least one treatment to help improve lung function. While existing research has studied the effects of individual treatments, the joint treatment effects are unknown. We applied the methods to estimate effects of two treatments (DNase and Hypertonic Saline) in combination on patients' lung function and use of IV antibiotics.

PARALLEL SESSION 25: Simulation and software

S25.1 Simple – A modular, open-source R software solution to SIMulate PLatform trials Efficiently

Elias Laurin Meyer[1], Tobias Mielke[2], Tom Parke[3], Franz Koenig[1]

[1] *Medical University of Vienna, Austria* , [2] *Janssen Cilag GmbH,* [3] *Berry Consultants*

Platform trials are becoming increasingly popular within drug development, attracting interest by patients, clinicians, regulatory agencies and certianly also statisticians [1]. More often than not, these platform trial designs are highly complex and involve too many weakly predictable events (e.g. number of investigational treatments that will enter over time) to determine the impact of relevant design parameters (e.g. decision rules, sharing of information across cohorts and allocation ratios) on the operating characteristics with high confidence. Simulations may address these uncertainties at the design stage. However, the number and combination of design elements for implementation in real platform trials is close to infinite. As a result, simulation software which is developed based on specific project needs is typically limited in the variety of available design options for comparison, as such software is developed for a particular need, not for researching all potential new approaches to clinical research and statistical science. On the other hand, software solutions, which allow for a wide range of design options, may easily overload the user with requirements for design specifications [2]. We have developed an R software package ("simple"), which is modular in the sense that if users want to simulate the most common platform designs, minimal effort and understanding of the package is needed, but it allows the users to take control of different parts of the simulation (e.g. patient accrual, outcome simulation, etc.) step-by-step, thereby facilitating the simulation of arbitrarily complex platform trials. We will give an overview of this software package alongside some examples on how to simulate the most common platform trial designs and derive their operating characteristics.

REFERENCES:
[1] Meyer, Elias Laurin, et al. «The evolution of master protocol clinical trial designs: A systematic literature review.» Clinical Therapeutics (2020). https://www.sciencedirect.com/science/article/pii/S0149291820302447
[2] Meyer, Elias Laurin, et al. "Systematic review of available software for multi-arm multi-stage and platform clinical trial design." Trials (2021). https://doi.org/10.1186/s13063-021-05130-x

S25.2 Simulating binary variables with two levels of clustering

Rhys Bowden[1], Jessica Kasza[1], Andrew Forbes[1]

[1] *Monash University, Melbourne, Australia*

Simulation is an important tool for assessing the performance of statistical methods. While many methods are available for the simulation of correlated binary random variables, all have significant practical limitations. These are due to computational infeasibility or because the range of allowable correlations with the method rapidly shrinks to zero as the number of random variables increases. In this paper we present a simple method for simulating multivariate binary random variables that is orders of magnitude faster than one of the commonly used simulation methods, while also allowing a much wider range of correlations than the other alternative methods.
This new simulation method is particularly suitable for the simulation of correlated binary data arising from longitudinal cluster randomised trial designs, such as the cluster randomised crossover trial and the stepped wedge trial designs. This simulation method allows for the outcome prevalence to change over time, and for a `nested exchangeable' correlation structure, in which observations measured in the same cluster in the same time period are more highly correlated than observations measured further apart in time. Further, this method is computationally efficient for simulating data for a large number of random variables (observations) per cluster. We have made the method available in an R package titled NestBin.

Parallel Sessions

S25.3 A collaborative approach to software development: The crmPack experience

Oliver Boix[1], Burak Kürsad Günhan[2], Thomas Jaki[3], Robert Adams[1], Phillip Crout[3], Pavel Mozgunov[3], Daniel Sabanes Bove[4], Christos Stylianou[5], Wojciech Wójciak[4], Dimitris Kontos[5]

[1] Bayer AG, Germany, [2] Merck KGaA, [3] MRC Biostatistics Unit, United Kingdom, [4] Roche, [5] ClinBay

Designing a phase I dose escalation trial using state of the art methodology often includes study simulations to derive operational characteristics. In addition, analysis software is needed to support dose escalation meetings during the conduct of the trial. Often in house developed software, proprietary/commercially available software or open source packages without reliable maintenance are used. Problems with this include the maintenance of in house software, the inflexibility of proprietary software and the lack of validation of open source software. Due to the continuous research and proposal of new methods to be used in phase I trials, the problem becomes even more pronounced. Recently, we evaluated the possibility to collaborate on the open source R package crmPack to overcome these problems, which was originally developed by Roche (https://CRAN.R-project.org/package=crmPack). crmPack provides a simple and unified object-oriented (S4) framework for model-based dose escalation designs. The package has already been used in some of phase I trials in the industry and academia individually - often tailoring it further to the specific needs. To avoid duplication and ensure continued maintenance, a group of industry, CRO and academia was put together for collaboration on the package. After a basic evaluation, we found that crmPack already covers a wide variety of methods and that different companies and institutions extended the package by including additional functions, more documentation, and testing for their needs separately. We concluded that instead of individual extensions, we can collaboratively further develop the package using modern software development methods and tools. Our current workflow takes place in Github (https://github.com/Roche/crmPack) to ensure version control and reliable collaboration. Furthermore, we are writing unit tests for the functions to prepare for subsequent validation of the package. In the talk, we will present the motivations, benefits, and the future plans of our experience.

S25.4 Simulation Guided Trial Design – The Challenges and The Benefits

Tom Parke[1], Dan Evans[2]

[1] Berry Consultants LLC, United Kingdom, [2] Pfizer UK

The design of more complex and innovative clinical trials requries the use of simulation of the trial design in order to understand and estimate the operating characteristics of the design. It turns out however that the process of simulating the design has a number of interesting and useful consequences over and above the ability to deliver an estimate of type-1 error and power for a design that othewise does not permit a closed form calculation. It also brings with it a number of interesting challenges that are statisticians are not normally primed to deal with.
We believe that clinical trial simulation is an enormously powerful tool, and not just for trials for which we can't calculcate the type-1 error and power. There are more things we should be exploring about our designs, such as the probability of selecting the right dose, the exepected time to decision, and the abilitiy to select the right patient populaiton. There are also aspects of conventional trials that frequently overlooked such as the effects of dropouts/missing data and time to the endpoint that can be explored with simulation.
An unexpected, less tangible, but sometimes crucial benefit of simulating the trial design, is how it can help communicate the properties of the design to the non-statistical members of the trial team. In particular by not just presenting overall results, but also presenting individual simulations.
However there are a number of new challenges the trial designer must overcome in order to prepare for simulating the proposed design, such as: gathering the additional assumptions needed to be able to simulate, selecting the scenarios to simulate, selecting and presenting the different operating charactistics carefully and clearly.
We hope that it would be helpful to present both these unexpected benefits, but also the unexpected challenges, to encourage statisticians to explore this powerful additional tool for trial design.

S25.5 Improving data transparency in the research community by constructing synthetic time-to-event data

Aiden Smith[1]

[1] University of Leicester, United Kingdom

A lack of availability of data and statistical code being published alongside journal articles provides a significant barrier to open scientific discourse, and reproducibility of research. Information governance restrictions inhibit the active dissemination of individual level data to accompany published manuscripts. Realistic, accurate time-to-event synthetic data can aid in the acceleration of methodological developments in survival analysis and beyond by enabling researchers to access and test published methods using data similar to that which they were developed on.
Here we propose methods to accurately emulate the covariate patterns and survival times found in real-world datasets using simulation techniques, without compromising individual patient identifiability. We model the joint covariate distribution of the original data using covariate specific sequential conditional regression models. A complex flexible parametric survival model is fitted from which survival times conditional on individual covariate patterns are simulated by drawing predictions from a centile distribution. This method is effective across both all-cause and competing risks settings, making it broadly implementable across all time-to-event data. We develop a series of metrics for evaluating the accuracy of the synthetic data and further show the ability to detect complex inter-relationships that were present in the real data. We further develop approaches to satisfy that individuals in the original dataset are not identifiable in the synthetic data created.
Using a series of case study datasets, including a freely available population level cancer registration dataset for colon cancer, we evaluate the effectiveness of the simulation methods for constructing synthetic data, by assessing covariate interrelationships and survival metrics from both an all-cause and competing risks perspective. We also provide evidence that it is almost impossible that a given patient from the original data could be identified from their individual unique date information. Simulated datasets using this methodology could be made available alongside published research without breaching data privacy protocols, and allow for data and code to be made available alongside methodological or applied manuscripts to greatly improve the transparency and accessibility of medical research.

PARALLEL SESSION 26: Missing data in Studies

S26.1 A flexible approach for the analysis of repeated attempt designs

Mike Daniels[1], Minji Lee[2]

[1] Department of Statistics, University of Florida, Gainesville FL, United States, [2] Edwards Lifesciences

It is not uncommon in follow-up studies to make multiple attempts to collect a measurement after baseline. Recording whether these attempts are successful or not provides useful information for the purposes of assessing the missing at random (MAR) assumption and facilitating missing not at random (MNAR) modeling. This is because measurements from subjects who provide this data after multiple failed attempts may differ from those who provide the measurement after fewer attempts. Previous models for these designs were parametric and/or did not allow sensitivity analysis. For the former, there are always concerns about model misspecification and for the latter, sensitivity analysis is essential when conducting inference in the presence of missing data. Here, we propose a new approach which minimizes issues with model misspecification by using Bayesian nonparametrics for the observed data distribution. We also introduce a novel approach for identification and sensitivity analysis. We re-analyze the repeated attempts data from a clinical trial involving patients with severe mental illness and conduct simulations to better understand the properties of our approach.

Parallel Sessions

S26.2 Non-parametric Multiple Imputation for Epoch-level Wearable data in Trials

Mia Tackney[1], Elizabeth Williamson[1], Derek G Cook[2], Elizabeth Limb[2], Tess Harris[2], James Carpenter[1]

[1] *London School of Hygiene and Tropical Medicine, United Kingdom*
[2] *St George's, University of London, United Kingdom*

Clinical trials that investigate the impact of an intervention on physical activity often use accelerometers to measure step count in very fine intervals of time, typically in 5-second epochs. The data is usually aggregated to provide daily or weekly step counts for the primary analysis. Missing data is common as participants may not wear the device as per protocol, or there may be device failure. Approaches to handling missing data in the literature have largely defined missingness on the day level using a threshold on wear time, which leads to loss of information on the time of day when data is missing. We present an approach to identifying and classifying missingness at the finer epoch-level. Missingness can then be handled using a non-parametric approach to Multiple Imputation (MI) where missing periods during the day are replaced by donor data from same person where possible, or data from a different person who is matched on demographic variables. The advantage of the non-parametric approach is that it is compatible with potentially complicated relationships in the accelerometer dataset, which do not have to be specified through a parametric model. We present results from simulation studies comparing the non-parametric MI to parametric MI, and illustrate the application of these different MI strategies to the analysis of the 2017 PACE-UP physical activity Trial.

S26.3 Targeting hypothetical estimands with causal inference and missing data estimators in a real trial

Camila Olarte Parra[1], Rhian Daniel[2], Jonathan Bartlett[1]

[1] *University of Bath, United Kingdom*
[2] *Cardiff University, United Kingdom*

In diabetes trials, rescue medication should be available for ethical reasons. Differential use of rescue medication between treatment arms may mask the effects of the study drugs. Estimating the treatment effect in the absence of rescue medication can then be of interest for certain stakeholders. Following the ICH E9 addendum terminology, the use of rescue medication would be an intercurrent event handled with a hypothetical strategy.
In this talk, we focus on the hypothetical strategy to deal with rescue medication and treatment discontinuation in a real diabetes trial. We illustrate how different estimators from the causal inference and missing data literatures can be applied to target different hypothetical estimands of interest. We also demonstrate how existing statistical packages can be exploited to estimate the different estimators. Finally, we discuss the challenges encountered in real life settings including multi-arm trials and missing data and provide potential solutions combining imputation methods and bootstrapping.

S26.4 Evaluation of multiple imputation approaches for case-cohort studies with binary outcomes

Melissa Middleton[1], Cattram Nguyen[1], Margarita Moreno-Betancur[1,2], John Carlin[1,2], Katherine Lee[1]

[1] *Murdoch Children's Research Institute, Melbourne, Australia,* [2] *University of Melbourne, Australia*

Case-cohort studies are a common variant of traditional cohort studies, wherein a random subcohort is selected and collection of exposure data is limited to this subcohort and all cases, termed the case-cohort sample, leading to a large proportion of information missing by design. While standard analysis of case-cohort studies uses inverse probability weighting (IPW) to address the missing data, little research has been conducted into addressing the additional challenge of data missing "by chance". Multiple imputation (MI) has been suggested as a potential approach, but it is unclear how best to tailor the imputation model so that it is compatible with a target analysis using IPW. An alternative is to use MI to handle both the missingness by chance and by design. While this has been studied in the context of a time-to-event outcome, it is unclear if this approach has value in the context of a binary outcome.
We therefore conducted a simulation and a case study to assess the performance of approaches for handling the missingness by chance and by design in a case-cohort analysis with a binary outcome. These included a solely MI approach, a solely IPW approach, a number of approaches combining MI and IPW applied to either the case-cohort sample or the full cohort, and a complete-case analysis (CCA) applying IPW to the records without missing data by chance. We considered a range of realistic simulated scenarios, varying the amount of missingness by chance and by design, the missing data mechanism, and the sample size, and illustrated these approaches in a case study using data from the Barwon Infant Study.
Results show that a combined approach is approximately unbiased for estimation of the exposure effect in large sample sizes and provides confidence intervals with close-to-nominal coverage, while a solely MI approach exhibited large biases and over-coverage. Precision was similar across all MI approaches and scenarios, with expected gains compared to a solely IPW or CCA approach. Our findings suggest that the combined approach might be the most optimal in these settings.

S26.5 Outcome variances after dropout as an indicator of dropout bias in randomized controlled trials

Audinga-Dea Hazewinkel[1], Kate Tilling[1], Kaitlin Wade[1], Tom Palmer[1]

[1] *University of Bristol, United Kingdom*

Randomized controlled trials (RCTs) are considered the gold standard for assessing the causal effect of an exposure on an outcome, but remain vulnerable to bias, for example due to missing data. When outcomes are missing not at random (MNAR), for example when dropout depends on the unobserved outcome value itself, estimates from complete case analysis (CCA) will be biased. There is no statistical test for distinguishing between outcomes missing at random (MAR) and MNAR, and current strategies rely on comparing dropout proportions and covariate distributions, using auxiliary information to assess the likelihood of dropout being associated with MNAR. We propose using the observed variance difference across treatment groups as a tool for assessing risk of MNAR bias.
In an RCT, at randomisation, the distributions of all covariates should be equal in the populations randomised to the intervention and control arms. Under the assumption of homogeneous treatment effects, the variance of the outcome will also be equal in the two populations over the course of follow-up. We show that under MAR dropout, the observed outcome variances, conditional on the variables included in the model, are equal across groups, while MNAR dropout may result in unequal variances. Consequently, unequal observed conditional group variances are an indicator of MNAR dropout and possible bias of the estimated treatment effect.
Heterogeneity of treatment effect affects the intervention group variance, and is another potential cause of observing different outcome variances. We show that, for longitudinal data, we can isolate the effect of MNAR outcome-dependent dropout by considering the variance difference at baseline in the same set of patients that are observed at final follow-up, and, in doing so, assess if the average treatment effect is likely to be biased. We illustrate our method in simulation and in an application using individual-level patient data from an RCT investigating the benefit of acupuncture treatment for patients with chronic headaches (ISRCTN96537534). We propose employing the variance difference as a MNAR bias assessment tool, which is easily implemented using standard available software, and has a straightforward interpretation of results.

Parallel Sessions

PARALLEL SESSION 27: Prediction modelling

S27.1 Effective sample size: a measure of individual uncertainty in predictions

Doranne Thomassen[1], Saskia le Cessie[1], Ewout W Steyerberg[1], Hans van Houwelingen[1]

[1] *Leiden University Medical Center, Netherlands*

When using a statistical model to predict a patient's outcome, we are usually uncertain that the model represents reality (ontological uncertainty), nor that we can perfectly predict the future at all (aleatory uncertainty)[1]. Even when assuming the model is correctly specified, there is epistemic uncertainty in predictions: the model is estimated after only partially observing the target population. Epistemic uncertainty typically differs between individual patients, as not all patient profiles are represented uniformly in model development. In that sense, expressing a patient's predicted risk as a number "out of 100 people like you"[2] fails to capture that their prediction may effectively be based on far fewer than 100 people. We aimed to define an intuitive measure of individual epistemic uncertainty in predictions.
For a given patient, the variance of a prediction can be equated to the variance of the sample mean outcome in n* hypothetical patients with the same covariate values. This hypothetical sample size n* can be interpreted as the effective sample size of similar patients in the data used to make the individualized prediction, given that the model is correct. For linear and logistic regression models, we derived analytical expressions for this effective sample size. In addition, we illustrate the concept in example datasets of patients with acute myocardial infarction.
The effective sample size n* may facilitate communication of uncertainty about predictions in clinical encounters. In model validation, the distribution of n* in a sample indicates for which patients predictions are more and less certain, and whether they might be too uncertain for some. This assessment could be a valuable addition to overall predictive accuracy measures such as calibration and discrimination. Finally, we discuss how n* can be used in model development to balance accuracy versus uncertainty of predictions. We propose the effective sample size as a clinically interpretable measure of uncertainty in individual predictions. Its implications should be explored further for the development, validation and clinical implementation of prediction models.

SHORT REFERENCES:
[1] Spiegelhalter, D. Annu Rev Stat Appl 4, 31-60 (2017)
[2] Kattan, M. W. et al. J Urology 170(5), 1792-1797 (2003)

S27.2 Automatic detection of glaucomatous neuropathy from fundus images

Lauren Coan[1], Gabriela Czanner[1], Silvester Czanner[1], Bryan Williams[2], Krishna Venkatesh[3], Rengaraj Venkatesh[3], Colin E Willoughby[4], Srinivasan Kavitha[3], Swati Upadhyaya[3]

[1] *Liverpool John Moores University, United Kingdom,* [2] *Lancaster University, United Kingdom,* [3] *Aravind Eye Hopsitals, India,* [4] *Ulster University, United Kingdom*

Glaucoma is the leading cause of irreversible blindness worldwide. Early detection is crucial, allowing timely intervention to prevent blindness. The cup-to-disc ratio (CDR) is a clinical parameter for glaucoma assessment obtained from the ratio between the boundaries of the optic disc and cup, which is observed in a fundus image of the optic nerve head (ONH). Yet manual assessments are subjective, time-consuming, and costly.
Automated glaucoma detection is increasingly researched via one-step Deep Learning (DL) black-box methods. There are also two-step approaches: step one provides cup and disc boundaries (segmented image) via a DL algorithm. The second step then uses a statistical/machine learning approach to utilise parameters extracted from the segmented images for glaucoma detection. We conducted a comprehensive literature review of the two-step approaches. Such methods provide inherent interpretability via their production of segmented images. Also, some methods report performance capabilities comparable to that of one-step DL, in online datasets. Our review findings highlight the potential benefits of two-step approaches whilst further research is required to advance and externally validate the methods. We extend the linear-mixed-effects model from MacCormick et al. (PLoSONE, 2019) to spatially model the shape of CDR profiles of the whole ONH (across 360 directions) whilst accounting for inter-subject differences via random effects. We fit the model using a combination of fixed and random effects including the main effect of groups (glaucoma and healthy), a collection of harmonic functions (to account for optic disc shape) and interactions effects. We use the fitted parameters to estimate the multivariate distributions of 360 CDR-profile vectors, and then calculate the posterior probability of glaucoma for each patient. Subsequently, a threshold is applied for glaucoma discrimination. The model is trained on the publicly available ORIGA database and tested on a retrospective study database from Aravind Eye Hospital, India. On external testing, the AUROC is 0.96, sensitivity 93% and specificity 92%. Our two-step approach to glaucoma detection allows us to explain the disease detection by highlighting optic disc deformation. Although our method shows the potential to give high accuracy, further validation via a prospective study is required.

S27.3 A comparison of methods for incorporating information from histtorical prediction models

Philip Boonstra[1], Pedro Orozco del Pino[1]

[1] *University of Michigan, Ann Arbor MI, United States*

When a new prediction model is developed, it it is often proposed in the context of existing prognostic models for the same outcome of interest. In limited-sample size scenarios, it is therefore desirable to try to borrow the information from these existing 'historical' models to make the proposed model more statistically efficient. However, these historical models may not be directly transportable to the current model of interest, for example, due to differences in the set of covariates adjusted for between each model or differences in the target population. To that end, several methods for data-integration have been proposed, each of which seeks to properly incorporate this historical information, including a semiparametric constrained maximum likelihood approach (Chatterjee, et al. (2016), JASA, 111:107-117), an extension that incorporates historical model uncertainty (Zhang, et al. (2020), Biometrika, 107:689-703), and an adaptive Bayesian shrinkage approach using the power prior (Boonstra and Barbaro (2020), Biostatistics, 21:e47-e64). Each method admits unique strengths and assumptions about the joint distribution of the covariates and the commensurability of the historical and current target populations. In the present work, we present a comprehensive simulation-based comparison of each of the aforementioned methods in a typical context of predicting a binary outcome. Our objective is to characterize scenarios in which a particular data-integration approach is appropriate and to assess the prediction and estimation performance of all methods. We also apply these methods to a real data analysis in which an existing mortality model for patients on ECMO is updated. In general, we find that the constrained maximum likelihood approach and its generalization perform better in scenarios with small-to-moderate numbers of predictors (e.g. 10-15) but that the algorithms for fitting these methods tend to encounter numerical difficulties related to a required matrix inversion in settings with larger numbers of covariates. In contrast, the adaptive Bayesian approach has statistical-bias-driven deficiencies in small-p settings but performs relatively much better in larger-p settings due to a favorable bias-variance tradeoff. Thus, we find that each approach has favorable characteristics, and our work points toward logical extensions for each of these approaches.

S27.4 Comparing methods to generate disease risk factor trajectories for longitudinal microsimulations

Christopher Jackson[1], Oliver Church[1]

[1] *MRC Biostatistics Unit, University of Cambridge, United Kingdom*

The long-term health impacts of interventions to prevent cardiovascular disease are often assessed by longitudinal microsimulation models. In these models, synthetic trajectories of risk factors through time are generated, alongside the resulting disease incidence and mortality. To assess the impacts of interventions, simulations are performed with risk factor trajectories modified in different ways under different interventions. This talk will compare different methods for generating risk factor trajectories for these models based on data, including previously-used and novel techniques.
One set of methods is based on jointly modelling multiple outcomes from longitudinal data. Correlations between successive measurements from the same person are represented either through a stochastic process (e.g. a Markov model) or through random effects. Parametric models will be compared with a nonparametric method based on regression trees.
Another set of methods uses cross-sectional data from populations of different ages. This is typically easier to obtain than comparable longitudinal data for the population of interest, and tends to cover a wider range of ages and time periods. Additional assumptions are required in this case, however, to simulate plausible synthetic longitudinal data. For example, we could assume that, as a person ages, they remain on the same quantile of the distribution of some risk factor relative to people of the same age.
We empirically compare the accuracy of simulated trajectories between different methods, in a case study of a model to assess the effectiveness of mid-life general health checks. This uses longitudinal data from the English Longitudinal Study of Ageing, and cross-sectional data from the Health Survey for England.

Parallel Sessions

S27.5 Dose-response prediction for in-vitro drug combination datasets: a probabilistic approach

Leiv Rønneberg[1], Paul DW Kirk[1], Manuela Zucknick[2]

[1] *MRC Biostatistics Unit, University of Cambridge, United Kingdom,* [2] *University of Oslo, Norway*

High-throughput drug sensitivity experiments in cancer enable rapid in-vitro testing of various compounds on cancer cell lines, or patient-derived material, in order to determine the efficacy of a certain treatment. Accurate prediction of dose-response functions from a limited set of pre-clinical experiments is key to explore the large space of possible treatment options, or to prioritise which experiments to perform. This is particularly important when predicting the effect of drug combinations, where it is unfeasible to test all possible combinations.
Drug sensitivity experiments are noisy by nature, due in part to the natural biological variability of cell growth but also technical error sources in the assays. This entails that the experimental observations of dose-response at different concentrations of a drug vary in estimation certainty both within and between experiments — a variability that has been often ignored in the literature.
We propose PIICM, a probabilistic framework for dose-response prediction in high-throughput drug combination datasets. PIICM utilizes a permutation invariant version of the intrinsic co-regionalization model for multi-output Gaussian Process regression to predict dose-response surfaces in untested drug combination experiments. The permutation invariance accounts for natural symmetries in the dose-response surfaces for drug combinations, which when not accounted for can have detrimental effects on prediction performance. Coupled with an observation model that incorporates experimental uncertainty, PIICM is able to learn from noisily observed cell-viability measurements in settings where the underlying dose-response experiments are of varying quality, and the training dataset is sparsely observed. We show that the model can accurately predict dose-response in held out experiments, and the resulting function captures relevant features indicating synergistic interaction between drugs.

PARALLEL SESSION 28: Efficient trial designs

S28.1 Two one-sided test-then-pool method for clinical trials

Kazufumi Okada[1], Isao Yokota[1]

[1] *Hokkaido University, Japan*

On designing a new clinical trial, through pooling the control arm data with historical data, there are merits in terms of more accurate estimates, increasing power, and reducing sample size. However, if there is heterogeneity between historical and current control populations, type I error rate may increase, or power may decrease. The test-then-pool method may combat these difficulties, which examines the similarity between the historical and current data using hypothesis testing, then it pools the historical data into the current data when there is no significant difference. A two-sided significance level of a test determines the degree of similarity between historical and current data for pooling, controls type I error rate and power simultaneously.
In this study, we extend the original test-then-pool method to two one-sided test-then-pool method, which decomposes one two-sided null hypothesis into two one-sided null hypotheses. Our method pools historical data into current data when both null hypotheses are not rejected. With two one-sided significance levels, our method can control type I error rate and power separately. Through a simulation study, we compare our method to the original method in terms of type I error rate and power. As a result, in our method, when there is a lack of similarity, the decrease of power becomes smaller than the original test-then-pool method under the same type I error rate. In the case when the power is equal to the original test-then-pool method, the pooling probability can increase instead.

S28.2 Using the Probability of Longer Survival to Assess the Efficacy of New Cancer Therapies

Michael LeBlanc[1], Yingqi Zhao[1], Mary Redman[1]

[1] *Fred Hutchinson Cancer Center, Seattle WA, United States*

Immunotherapies and targeted therapies have brought meaningful improvements in survival for patients but also newer challenges to the design and interpretation of clinical trials. As ubiquitous is the use of hazard ratios to summarize treatment effects for time-to event outcomes, the hazard ratio is challenging to explain to non-statisticians. Further the hazard ratio does not summarize treatment effects well when treatments have delayed effects or when there is heterogeneity in treatment efficacy which can happen with biomarkers determining differential activity or cure-fractions.
We investigate alternative measures to the hazard ratio to summarize treatment effects. The probability of longer survival (PoLS), or variations on this quantity (Peron et al, 2015) are an attractive and intuitive way to summarize benefits of an investigational therapy. The PoLS quantifies the treatment effects as probability a random patient on an investigational treatment lives an additional time m (where m could be months) compared to a random patient on standard therapy for, providing a clinically interpretable quantification of important treatment effects. Additionally, parameters based on the PoLS are interpretable when there are delayed treatment effects, crossing curves, or cure fractions, all of which result in a violation of the proportional hazards assumption and challenge the interpretability of the hazard ratio.
We build on our prior work Zhao et al. Stat in Med, 2019, on PoLS, to calculate the adjusted probability of longer survival, (APLS), and the relative probability of longer survival (RPLS) and show how they can be estimated both from raw clinical trials data and also using published survival curves. Further, to represent these parameters, we develop and describe the Probability of Longer Survival Area (PoLSA) plot which utilizes flexible parametric estimates of the survival distributions. Data from both simulated trials and completed trials are used to demonstrate the properties of the PoLSA plots. We include results using data from SWOG Cancer Research Network (part of the US National Cancer Institute (NCI), National Clinical Trials Network) as well data extracted from other published positive immunotherapy trials.

S28.3 Testing for treatment differences with allocation probabilities in response adaptive trials

Thomas Jaki[1], Helen Barnett[1]

[1] *MRC Biostatistics Unit, United Kingdom*

Commonly the objective for response adaptive clinical trials is to ensure that patients within a trial have a high chance of receiving the best treatment available by altering the chance of allocation on the basis of accumulating data. Approaches which yield good patient benefit properties suffer from low power from a frequentist perspective when testing for a treatment difference at the end of the study due to the high imbalance in treatment allocations.
In this work we develop an alternative pairwise test for treatment difference on the basis of allocation probabilities of covariate-adjusted response-adaptive randomization with forward looking Gittins index rule (CARA-FLGI). The performance of the novel test is evaluated in simulations for two-armed studies and then its applications to multi-armed studies is illustrated.

28.4 SDetermining sample size in a personalised randomised trial comparing a network of treatments

Becky Turner[1], Sarah Walker[1], Sally Ellis[2], Mike Sharland[3], Ian White[1]

[1] MRC Clinical Trials Unit at UCL, United Kingdom, [2] Global Antibiotic R&D Partnership (GARDP), [3] St. George's, University of London, United Kingdom

In some clinical settings, multiple treatment regimens are potentially useful for treating the disease of interest, but some treatments may not be appropriate for certain subgroups of patients. A personalised randomised controlled trial (PRACTical) design has been proposed for this setting. For a network of treatments, each patient is randomised only among treatments for which they are eligible. The aim is to produce treatment rankings that can inform clinical decisions about treatment choices for individual patients. Here we propose methods for determining sample size in a PRACTical design, since standard power-based methods are not applicable when comparing a network of treatments.
We derived a sample size by evaluating how much information would be gained from trials of varying sizes. For a binary outcome, we considered how many adverse outcomes would be prevented by choosing the top-ranked treatment for each patient based on the trial's results rather than choosing a random treatment from the appropriate personalised randomisation list before the trial. In a simulation study, we evaluated three performance measures: average reduction in adverse outcomes using sample information, proportion of simulated trials in which the top-ranked treatment had adverse outcome frequency within 1% of the best treatment, and proportion of simulated trials in which the top-ranked treatment was better than a randomly chosen treatment.
Our example is a trial evaluating eight different combination antibiotic regimens for neonatal sepsis (NeoSep1 Trial), in which a PRACTical design addressed varying patterns of antibiotic resistance and clinical acceptability. A second PRACTical randomisation to second-line treatment was allowed where required. In a simulation study, we evaluated trial performance with respect to ranking first-line and second-line treatment strategies across several plausible scenarios. Using top-ranked treatment strategies from a trial including 3000 babies was estimated to achieve at least 80% of the maximum possible reduction in mortality, with 80% chance of selecting a treatment within 1% of the best treatment strategy and 97% chance of being better than a random strategy. Calculating the value of treatment ranking information was an effective approach for determining sample size in a PRACTical trial comparing a network of treatments.

S28.5 Solutions for Surrogacy Validation with Longitudinal Outcomes for a Gene Therapy

Emily Roberts[1], Michael Elliott[1], Jeremy Taylor[1]

[1] University of Michigan, Ann Arbor MI, United States

Valid surrogate endpoints S can be used as a substitute for a true outcome of interest T to measure treatment efficacy in a clinical trial. We propose a causal inference approach to validate a surrogate by incorporating longitudinal measurements of the true outcomes using a mixed modeling approach, and we define models and quantities for validation that may vary across the study period using principal surrogacy criteria. We consider a surrogate-dependent treatment efficacy curve that allows us to validate the surrogate at different time points. We extend these methods to accommodate a crossover design where all patients eventually receive the treatment. Because not all parameters are identified in the general setting, we rely on informative prior distributions to obtain inference. We consider the sensitivity of these prior assumptions as well as assumptions of independence among certain counterfactual quantities conditional on pre-treatment covariates to improve identifiability. We examine the frequentist properties (bias of point and variance estimates, credible interval coverage) of a Bayesian imputation method. Our work is motivated by a clinical trial of a gene therapy where the functional outcomes are measured repeatedly throughout the trial.

PARALLEL SESSION 29: Joint inference with high dimensional data

S29.1 Bayesian variable selection with applications to selection of comorbidities

F Javier Rubio[1], David Rossell[2], Miguel Angel Luque-Fernandez[3]

[1] University College London, United Kingdom, [2] Pompeu Fabra University, Barcelona, Spain, [3] London School of Hygiene and Tropical Medicine, United Kingdom

In the first part of the talk, we will discuss the role of misspecification and censoring on Bayesian model and variable selection in the context of right-censored survival regression. Misspecification includes wrongly assuming the censoring mechanism to be non-informative and/or incorrectly specifying the survival model. Emphasis is placed on additive accelerated failure time and Cox proportional hazards with local and non-local priors. We discuss a fundamental question: what solution can one hope to obtain when (inevitably) models are misspecified, and how to interpret it? We show that misspecification and censoring have an asymptotically negligible effect on false positives, but their impact on power is exponential. In the second part of the talk, we apply this methodology to identify the comorbidities that predict overall survival in lung and colorectal cancer patients. The aim of this study is to understand how different cancer sites and tumour stages are affected by different comorbidities. Identifying the comorbidities that affect cancer survival is indeed of interest as this information can be used to identify factors driving the survival of cancer patients at the population level. The proposed methodology for Bayesian variable selection is available in the R package 'mombf'.

S29.2 A scalable ECM algorithm for multiple-network joint inference with the graphical horseshoe

Camilla Lingjaerde[1], Hélène Ruffieux[1], Sylvia Richardson[1]

[1] MRC Biostatistics Unit, University of Cambridge, United Kingdom

In statistical omics, network models are useful tools for modelling complex associations and assessing pathway activity. If a Gaussian graphical model is assumed, an association network can be estimated by determining the non-zero entries of the inverse covariance (precision) matrix of the data. Within a Bayesian framework, the graphical horseshoe estimator provides a robust and flexible framework for precision matrix inference in Gaussian graphical models, as it assumes local, edge-specific parameters which prevent overshrinkage of non-zero off diagonal elements. However, in the high-dimensional settings commonly found in biological networks, the Gibbs sampler for the graphical horseshoe is often computationally inefficient or even impractical. Further, the model is only formulated for a single network and does not offer an integrative analysis should there be multiple related data sets available that might share common structures. We propose a novel scalable expectation conditional maximization (ECM) algorithm for obtaining the posterior mode of the precision matrix in the graphical horseshoe. Moreover, we extend our framework to a multiple network setting, so that edge-specific information can be shared between related networks to improve estimation. We apply our method to both simulated and real omics data sets, and show that our approach outperforms the existing Gibbs sampler both in terms of computability and accuracy. We also show that the joint network approach successfully shares information between networks while capturing their differences, ensuring that information is shared only to the degree that it improves the model fit and ensuring better interpretability of the resulting edges.

S29.3 A Natural History and Copula Based Joint Model for Regional and Distant Breast Cancer Metastasis

Alessandro Gasparini[1], Keith Humphreys[1]

[1] *Karolinska Institutet, Sweden*

The clinical staging of breast cancer is strongly correlated with prognosis and contributes to determining specific treatment regimes. In Sweden, between 1992 and 2012, the 5-year relative survival rates ranged between 97-98%, 92-94%, 75-79%, and 25-33% for women diagnosed with Stage I to IV breast cancer, respectively. Despite the importance of staging for prognosis, there are only a few statistical models of breast cancer recurrence and progression to distant metastasis, with most approaches based on multi-state modelling. While these are very useful for summarising the risk of recurrence, they can become very complex and are not based on underlying biological mechanisms. They have limited use for understanding the implications of interventions at a population level (such as novel screening strategies). We have developed an alternative, flexible, and parsimonious modelling approach for latent tumour growth and spread to local and distant metastasis, based on a natural history model with biologically inspired components. Our procedure includes marginal submodels for local and distant breast cancer metastasis, which we jointly model using a copula function; different copula formulations (and correlation shapes between the margins) are allowed and implemented. In this way, we can incorporate and directly model the correlation between local and distant metastasis flexibly and efficiently. Sub-models for the latent growth of breast cancer, the detection process, and screening sensitivity, together with random effects to account for between-patients heterogeneity in tumour growth and spread, are also included in the joint model. The resulting joint likelihood function can be maximised using any general-purpose optimiser.
Although relying on a variety of parametric assumptions, the joint copula model can be useful for (1) understanding – potentially latent – disease dynamics, (2) obtaining patient-specific, model-based predictions, and (3) studying interventions at a population level, e.g., within a microsimulation framework.
We illustrate our approach by analysing data from a Swedish population-based case-control study of postmenopausal breast cancer. We also provide examples of model-based predictions.

S29.4 Marker selection in joint analysis with competing risks: application to SARS-COV-2 patients.

Alexandra Lavalley-Morelle[1], Xavier Lescure[1], Nathan Peiffer-Smadja[1], Simon Gressens[2], Alexandre Lahens[2], Agathe-Julie Bounhol[2], France Mentré[3], Jimmy Mullaert[1]

[1] *Université Paris Cité, IAME, Inserm, France,* [2] *Department of Infectious and Tropical Diseases, AP-HP, Bichat-Claude Bernard University Hospital, France,* [3] *Inserm UMR 1137 IAME*

During the Covid-19 pandemic, a large number of clinical prognostic scores have been proposed and evaluated for hospitalized patients, exclusively relying on variables available at admission. However, capturing data collected from the longitudinal follow-up of patients during hospitalization may greatly improve prediction accuracy. The joint modeling approach provides individual dynamic predictions that take the full history of longitudinal measurements into account. The objective is to assess the added value of integrating routine biological follow-up with respect to existing baseline scores, in order to predict the risk of in-hospital death.
Patients diagnosed with Covid-19 and hospitalized in the infectious disease department of Bichat hospital (Paris, France) between January and July 2020 were considered for the analysis. The primary outcome was time until in-hospital death. Discharge was treated as a competing event. All the results of available biological exams were collected. Each marker is modelled by parametric linear or non-linear mixed-effects model jointly estimated with a parametric competing risks model involving subdistribution hazards. The survival model is adjusted on the value of the 4C score and involves a coefficient linking the current value of the biomarker to the instantaneous subdistribution hazard. Estimation is performed on Monolix software with SAEM algorithm. After goodness of fit criteria, the link coefficient is tested for each marker using Wald test. Multivariable analysis is performed. We derive time-dependent AUC, scaled Brier score, and their 95% confidence intervals.
After quality control, 327 patients and 59 biomarkers are considered in the analysis. The baseline score, including age, gender, comorbidities, and baseline urea and CRP already has a strong effect on in-hospital death (exp(β)=1.43 95%CI [1.30;1.58]). After joint model estimation, 26 biomarkers are significantly associated with the score-adjusted instantaneous risk of death. Multivariable analysis will be presented.
The originality of the work relies on (i) the statistical approach combining a (possibly non-linear) longitudinal model jointly estimated with a parametric subdistribution model to derive individual dynamic predictions, and (ii) the use of real-life hospital data to collect massive data on biological examinations.

S29.5 Modelling dynamic associations in multivariate longitudinal data

Roula Tsonaka[1], Georgy Gomon[1], Bart Mertens[1], Dimitris Rizopoulos[2]

[1] *Leiden University Medical Center, Netherlands,* [2] *Erasmus Medical Center, Rotterdam, Netherlands*

In longitudinal studies measuring the association between the longitudinal response variable and time-varying covariates such as, the treatment dose, environmental exposures or biomarkers can be challenging for a number of reasons. First, the time-varying covariate may be endogenous and thus standard multivariate models e.g., mixed effects models for the longitudinal outcome will be biased unless the covariate process is explicitly modeled. Second, the outcome and the covariate process may not be measured at the same timepoints or missing data in any of the two processes can complicate their joint analysis. Finally, the functional form of the association is often not known beforehand and thus several options need to investigated. To address these challenges we will review several approaches that differ on the assumed association structure between the involved processes. In particular, we consider multivariate models that induce the association via correlated random effects and joint models that factorize the joint distribution of the two processes in terms of conditional densities. Nevertheless, fitting these models is not straightforward and their computational intensity due to the potentially high-dimensional integrations over the random effects terms limits their applicability. Therefore, in this work we propose to fill this gap by using nested Laplace methods as implemented in the R package INLA. This is a flexible Bayesian estimation approach which avoids the need for MCMC approximation and has been successfully and efficiently applied for the estimation of a large class of Bayesian statistical models relevant to biostatistics applications. To illustrate the feasibility of the proposed approach we will present analysis results from application on clinical studies at our home University Medical Center.

PARALLEL SESSION 30: Machine learning and prediction

S30.1 Deep Survival EWAS modeling of cancer time to diagnosis via blood-derived DNA methylation profiles

Francesca Ieva[1], Michela Massi[1], Giovanni Fiorito[2]

[1] *Politecnico di Milano, MOX Department of Mathematics, Italy,* [2] *Università di Sassari, Italy*

DNA methylation (DNAm) on CpG islands is an epigenetic process that was proven to be implicated in cancer. Therefore, blood-derived DNAm profiles could become biomarkers for cancer risk stratification and early detection. Moreover, pre-diagnostic data opens the further opportunity to investigate the relationship between sites' methylation at baseline and the Time to Diagnosis (TTD). Previous studies provide inconsistent results, due to the limitations of standard Epigenome-wide association studies (EWAS) approaches, that model each island's effect independently.
Here we argue that a novel global approach that captures the complex (potentially non-linear) relationships interplaying between sites is needed. Therefore, we develop a new Deep Learning (DL)-based approach assessing the relevance of individual CpG Islands (i.e., assigning a weight each) in determining TTD while modeling their combined effect in a survival analysis scenario.
Our Deep Survival EWAS approach is composed of several steps, that are meant to deal with the common complexities of epidemiological studies, such as small sample size and huge dimensionality, noise, and low signal-to-noise ratio of blood-derived DNAm. Indeed, it first alleviates noise, dimensionality and collinearity by aggregating CpG Islands via hierarchical clustering. Then, a tailored sampling procedure generates K independent training-test sets to foster robustness in the derived weights profile. Indeed, each of the K training sets is independently used to train a Deep Survival Model to predict TTD, while the relative test set is supplied to Shapley Additive Explanation (SHAP) algorithm, that explains the model by associating a weight to each feature in terms of importance in determining the prediction. The K sets of weights profiles are aggregated to obtain the final estimation of the effects profile, that is redistributed to single CpG Islands assigning each site the value of their corresponding feature cluster.
We apply our approach to a prospective case-control study on breast cancer nested in the EPIC Italy cohort and a weighted gene set enrichment analysis confirms that it outperforms standard EWAS in identifying biologically meaningful pathways. Moreover, we confirm the value of a DL-based model accounting for predictors interactions comparing our method with a Cox model with additive effects only.

Parallel Sessions

S30.2 A Comprehensive Framework for the Evaluation of Individual Treatment Rules From Observational Data

Francois Grolleau[1], François Petit[1], Raphaël Porcher[1]

[1] *Centre of Research in Epidemiology and Statistics Sorbonne Paris Cité, France*

INTRODUCTION: Individualized treatment rules (ITRs) are deterministic maps that recommend a treatment option to individuals based on their characteristics (we consider here the case two treatment options only). Randomized trials that evaluate the use of an ITR vs usual care are rare.
We introduce a method to emulate such trials from observational data distinguishing between two situations: i) the situation of a newly developed ITR, where data are from a population where no patient implements the ITR, and ii) the situation of a partially implemented ITR, where data are from a population where the ITR is stochastically implemented in some a priori unidentified patients.
METHODS: Our main estimand, the Average Rule Effect (ARE), represents the population-level effect of an ITR on an outcome in a trial where patients would have been randomized to receive either the treatment recommended by ITR or usual care.
In the new ITR situation, we rely on Rubin's causal model to adapt classical estimators and derive a double robust estimator for the ARE. In the partially implemented ITR situation, on top of the fundamental problem of causal inference, we face the challenge of unknown stochastic implementation functions. For this situation, we introduce a new causal model and propose a procedure to estimate the ARE using a hierarchical mixture of experts fitted via an expectation-maximization algorithm.
RESULTS: To investigate the properties of our ARE estimator in the partially implemented ITR situation, we conduct Monte Carlo simulations where we vary the sample size and data generating process. Our ARE estimator exhibits decreasing bias and variance as the sample size increases. Bias appears always toward the null and lower when treatment allocation in the absence of the ITR resembles the ITR.
We illustrate our approach on the MIMIC-III electronic health record, focusing on ITRs for initiation of renal replacement therapy in patients with acute kidney injury. We evaluate two ITRs that either i) recommends renal replacement therapy initiation based on six biomarkers (new ITR situation) or ii) relies on Sequential Organ Failure Assessment (SOFA) scores for such recommendations (partially implemented ITR situation).

S30.3 Quantifying instability after developing a clinical prediction model

Richard Riley[1], Gary Collins[2]

[1] *Keele University, Newcastle, United Kingdom,* [2] *University of Oxford, United Kingdom*

BACKGROUND: Clinical prediction models estimate an individual's risk of an outcome being present (diagnostic model) or occurring in the future (prognostic model), conditional on their values of multiple predictors. A developed model is a consequence of the sample used to develop it and the chosen modelling strategy, including the predictors considered and analysis approach taken (e.g., regression, machine learning; penalization methods; handling of missing values). A concern is that, in practice, many models are developed using weak analysis methods and small datasets, such that model predictions are unreliable in new data. We refer to this problem as model instability.
OBJECTIVES: The talk aims to: (i) show how instability is a consequence of sampling variability in the context of the chosen model development approach and sample size of the development dataset; (ii) demonstrate how instability manifests itself as miscalibration of model predictions in new data, and (iii) provide recommendations for how to explore and quantify instability after a model is developed.
METHODS: Simulation and case studies of instability, for a variety of statistical and machine learning methods for model development.
RESULTS: Simulation shows that instability of model predictions is often considerable, regardless of the modelling techniques chosen, even when sample size seems reasonable. Some individuals' predictions may even vary from 0 to 1. In practice, researchers should use bootstrapping to assess instability. This entails redeveloping their model (using the same original model development steps) in each bootstrap sample, and presenting the variability of predicted risks on a "prediction instability plot" of bootstrap model predictions (y-axis) versus original model prediction (x-axis). Instability can also be quantified using the instability index (the mean absolute difference between the original and the bootstrap model predictions) and via a "calibration instability plot" displaying the variability of calibration curves for the original model's predictions in the bootstrap samples.
CONCLUSION: Instability measures and plots should be routinely presented in studies that develop a prediction model. They add important information over other summary measures (e.g. C-statistics and calibration plots), and expose whether the model is likely to be unreliable.

S30.4 History-Restricted MSM and LCGM of Treatment Trajectories for a time-dependent outcome

Awa Diop[1], Caroline Sirois[1], Denis Talbot[1]

[1] *Laval University, Quebec QC, Canada*

Prevention of cardiovascular diseases (CVDs) is crucial given their high frequency and impact on population health. Statins, medications that reduce cholesterol levels, are used in primary prevention of CVDs, but their benefits for this purpose among older individuals are uncertain. In this study, our goal is to estimate the effect of time-varying statin usage on the time to a first CVD event using medico-administrative databases. A widespread approach to measure the impact of statins on CVDs is by estimating the effect of the cumulative number of periods that subjects were exposed to the treatment. Alternatively, some researchers are more interested in knowing the effect of treatment adherence often measured using medication possession ratio or proportion of days covered. However, these methods do not capture the complex dynamics of a time-varying treatment. Latent Class Growth Model (LCGM) are increasingly adopted as a better method to measure adherence. In a previous work, we introduced a framework that combines LCGM with marginal structural models (LCGM-MSM). LCGM-MSM first summarizes the numerous time-varying treatment patterns into a few trajectory groups and then allows for a population-level causal interpretation of group effects on the outcome. Despite these advantages, the LCGM-MSM framework is not suitable when the outcome is time-dependent. We propose combining a nonparametric history-restricted MSM (HRMSM) with LCGM. HRMSMs consist of defining a shorter history of exposure and are seen as a repeated application of standard MSMs. To the best of our knowledge, we present the first application of HRMSMs with a time-to-event outcome. It was previously noted that HRMSMs could pose interpretation problems in survival analysis. We propose a causal parameter that bypasses these interpretation challenges. We consider three different estimators of the parameters: inverse probability of treatment weighting (IPTW), g-computation, and a pooled longitudinal targeted maximum likelihood estimator (pooled LTMLE). We conducted simulation studies to measure the performance of the proposed LCGM-HRMSM. For all scenarios, we obtained unbiased estimates when using either g-computation or pooled LTMLE/pooled LTMLE + Super Learner. IPTW produced estimates with slightly larger bias in some scenarios. All approaches had good coverage of the 95% confidence interval.

S30.5 Flexible parametrization of individual sparsified networks for prediction: a proof-of-concept

Mariella Gregorich[1], Georg Heinze[1]

[1] *Medical University of Vienna, Austria*

Important aspects of the development of prediction models are the identification of potential predictors of outcome and the specification of an appropriate model structure to accurately capture their association with the outcome. Nowadays, many structurally similar predictors are available for each individual, which can be represented as individual-specific networks which capture their dependence structure and provide predictive features for outcome modelling. However, unsubstantiated, arbitrary decisions in individual-specific network inference, in particular when choosing a suitable threshold for sparsification, can lead to a high variability of the extracted graph-theoretical features. We propose a flexible parametrization approach to include graph-theoretical features as explanatory variables in a prediction model. In particular, flexible functional weight functions of the threshold value determined by statistical goodness-of-fit criteria enable us to incorporate uncertainties in network inference in the model. We perform a simulation study to provide evidence for a proof-of-concept of our proposed methodology in individual-specific networks of a given size, density and with particular, well-defined network properties. We compare the predictive performance of our approach to a more conventional method of selecting a single sparsification threshold based on AIC. The estimands in the simulation study are the functional relationship of the network feature-outcome association and the predictive performance of models. Performance is assessed by the bias of the identified maximum of the estimated weight function compared to the true maximum, the root mean squared error function for the estimated weight function and the average root mean squared error of the predictions in an independent validation set.
The new flexible approach can provide noticeable advantages compared to previous approaches. However, extensive simulation studies are necessary to assess the impact and ambiguity of common network inference approaches in clinical data sets. We highlight some challenges that need to be addressed before our approach is ready for routing applications and provide recommendations of our proposed approach in an applied data setting.

PARALLEL SESSION 31: Early Phase Trials

S31.1 Decision making under uncertainty in PI-II dose finding trials in Oncology

Andrew Hall[1], Duncan Wilson[1], Stuart Barber[1], Sarah Brown[1]

[1] *University of Leeds, United Kingdom*

There is increased interest in dose finding methods in oncology using both toxicity and efficacy endpoints with targeted therapies. Patients typically seek treatment benefit from entering such a trial. It is therefore unethical to put patients at undue risk of toxicity or treat at doses when evidence suggests alternatives may be more beneficial. A phase I trial design proceeds in stages with a decision as which dose to give the next group of patients made after every stage. The success of a trial is measured by its ability to locate an optimal dose and its tendency to suggest treatment at undesirable doses. Bayesian decision theoretical approaches have previously been found to be in theory ethically and scientifically sound. In practice however, it is challenging to specify a utility function capturing clinical preferences while maintaining good operating characteristics sensitive to utility specification.
Outcomes from treatments are not deterministic; utilities are a measure of preference when facing an uncertain outcome. We consider situations where preferences/utilities are defined for binary efficacy and toxicity attributes with respect to outcome probabilities. In doing so, clinicians can account for individual patient risk while meeting wider trial objectives, i.e. identifying a recommended phase II dose. The bivariate utility is formed by inspecting utility independence axioms to give a simplified linear structure, allowing us to concentrate on more easily assessed univariate utility functions. We argue attitudes to risk for univariate utility functions follow heuristics from prospect theory. Namely they are framed from the perspective of a reference point, with a risk averse attitude for perceived gains, and risk seeking for losses. Additionally, with loss aversion it is ethical to avoid losses more so than to pursue gains.
I will explain why heuristics from prospect theory, used to structure utilities around outcome probabilities, are justified in this setting. The method is compared for a range of scenarios to a simplified design with utilities based upon the four elementary outcomes from an individual patient. The design in general gives comparable operating characteristics but excels in more difficult scenarios when the optimal dose and/or unsuitable doses are close to stopping boundaries.

S31.2 Beyond 3+3 – acceptability and implementation of model-based dose-finding study designs in practice

Helene Thygesen[1]

[1] *University of Glasgow, United Kingdom*

Phase I trial protocols are usually based on mechanistic designs such as 3+3 or Rolling 6. There are several problems with such designs. The dose escalation process is inflexible, and the lack of modelling of the dose-response relationship makes it unclear how the data should be analysed. Several model-based approaches to phase I trial designs, such as the Continual Reassessment Method (CRM), have therefore been developed.
However, it can be a challenging process to get model-based designs accepted and implemented in practice. Formalizing modelling assumptions and prior beliefs is particularly difficult in early phase studies. And with the need for supporting dynamic decision making, simple and mechanistic dose escalation rules are attractive.
In this talk, I share some experience I have had with eliciting parameter values for Bayesian Nonparametric Methods and CRM from clinical investigators, and I present some methods that can be used to make the model-based approaches more transparent and easier to implement.

S31.3 Advanced tumor metrics to support characterization of the dose-response relationship

Cornelia Ursula Kunz[1], Katharina Piedboeuf[1], Stephan Lücke[1]

[1] *Boehringer Ingelheim Pharma GmbH & Co. KG*

Development of new drugs is a time-consuming and sometimes inefficient process. Especially in oncology, clinical trials tend to be expensive. Hence, there is a need to save time and cost and speed up the development process. Initiatives like the FDA's Critical Path Initiative aim at transforming the way FDA-regulated products are developed in general while the Oncology Center of Excellence Project Optimus specifically focuses on optimizing the dose selection process in oncology.
Typically, early phase trials in oncology rely on binary outcomes like objective response to evaluate anti-tumor activity, with the drug entering late phase development if the response rate achieves some pre-defined criteria. Unlike in other therapeutic areas, no formal dose finding is performed in the sense of establishing a dose-response relationship and selecting an optimal recommended Phase 2 dose (RP2D). One main question is whether advanced tumor-derived metrics can be used in addition to PK/PD modeling to better characterize the dose-response relationships and support RP2D definition.
Tumor growth models which describe the change of the tumor burden over time (in response to therapy) using exponential functions could provide alternative measures, as for example the modeling approaches which quantitatively capture tumor response to treatment using the g(rowth)-rate and d(ecline)-parameter (e.g. Wilkerson2017).
Based on the exponential tumor growth models, we investigate the mathematical properties of the models and derive several equations and algorithms linking the g- and d-parameters to other measures like (time-to)-response, -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. Using these constrains leads to unique solutions for the mean of the logarithm of the g- and d-parameter. The framework can be used to jointly simulate response and time-to-event endpoints in oncology. Based on this we 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.
REFERENCE: Wilkerson, J. et al. (2017). "Estimation of tumour regression and growth rates during treatment in patients with advanced prostate cancer: a retroperspective analysis". In: Lancet Oncol 18, pp. 143–154

S31.4 Improving interim decision in two-stage phase II designs by incorporating short-term endpoints

Dario Zocholl[1], Cornelia Ursula Kunz[2], Geraldine Rauch[1]

[1] *Charité Universitätsmedizin Berlin, Germany,* [2] *Boehringer Ingelheim Pharma GmbH & Co. KG*

In oncology, phase II clinical trials are often conducted to assess whether a treatment shows promising effect in patients that warrants further investigation of the treatment. Many trials include an option to stop for futility. For single-arm trials, the most popular approach is the Simon's two-stage design. However, when the time point of outcome assessment is long after the patient'sstudy entry, e.g. when primary endpoint is disease progression after 12 months, it may take an undesirably long to obtain the interim decision whether to stop for futility. Approaches that use short-term endpoints for futility stopping have been proposed but they require a priori assumptions about these short-term endpoints and about the correlation with the long-term endpoint. We show that with misspecified assumptions on the short-term endpoint, type I error rate and statistical power can be heavily affected. When information to make reliable assumptions is not available in the planning stage, approaches would be preferred that require only assumptions on the primary endpoint but can still incorporate information from short-term endpoints. We propose two approaches which could be applied in these situations, one of which is based on conditional power and the other is based on Bayesian posterior predictive probability of success. We present a calibration method to adhere to a pre-specified type I error rate, and for the Bayesian approach a weakly informative prior. We use simulations to show the performance of the two methods in comparison with existing approaches. We show that both methods provide stable operating characteristics across a wide range of scenarios if patient recruitment is rather slow and can considerably improve power and type I error rate compared to previous approaches.

S31.5 Aggregating prior distributions from experts for sample size calculations

Kevin Wilson[1], Nina Wilson[1], Cameron Williams[1]

[1] *Newcastle University, United Kingdom*

In order to choose a sample size for a clinical trial using elicited prior distributions from experts, it is first necessary to combine the individual priors into a single group prior. There are many methods of prior aggregation, which can roughly be categorised into two types. Mathematical aggregation methods combine prior distributions using a mathematical rule, while behavioural aggregation methods assist the group of experts to come to a consensus prior through discussion. As many commonly used aggregation methods have different requirements in the elicitation stage, there are few, if any, comparisons between them.
In order to choose a sample size for a randomised controlled trial into a novel diagnostic test for Motor Neuron Disease, we elicited a number of prior distributions from a group of experts. We then aggregated these prior distributions using a range of mathematical aggregation methods, including Equal Weights linear pooling, the Classical Method, and a Bayesian aggregation method. We also undertook an in-person behavioural aggregation with the experts, using the Sheffield Elicitation Framework, or SHELF.
In this presentation, using expert answers to seed questions, for which the elicitors know the true values, we compare and contrast the different aggregation methods and their performance. We also demonstrate how all considered aggregation methods outperform the individual experts. Finally, we compare the samples sizes calculated based on the different aggregation methods.

PARALLEL SESSION 32: Complex Modelling

S32.1 Flexible parametric methods for calculating life expectancy in small populations

Freya Tyrer[1], Nina Wilson[1], Yogini Chudasama[2], Paul Lambert[3,4], Mark Rutherford[5],

[1] *Biostatistics Research Group, Health Sciences, University of Leicester, United Kingdom,* [2] *Leicester Real World Evidence Unit, Diabetes Centre, Health Sciences, University of Leicester, United Kingdom,* [3] *University of Leicester, United Kingdom,* [4] *Karolinska Institute, Solna, Sweden*

Life expectancy is a simple measure to compare health differences between populations, but existing methods are not reliable for small populations. A potential solution to this problem is to 'borrow strength' from larger sub-populations within the same data set, but this has not previously been investigated. For this study, we used linked data from the Clinical Practice Research Datalink (CPRD) and Hospital Episode Statistics (HES) in England to compare stratified and combined flexible parametric models for calculating life expectancy by presence/absence of intellectual disability and type 2 diabetes mellitus (T2DM). Confidence intervals were calculated using the Delta method and bootstrapping. The Chiang's adjusted life table approach was used for comparison.
We found that the flexible parametric models had advantages over traditional methods by allowing calculation of life expectancy by exact age and beyond traditional life expectancy age thresholds, without the need to truncate age groups. The combined model that fit age interaction effects as a spline term provided greater statistical precision for the small covariate subgroups by borrowing strength from the larger subgroups. However, careful consideration was needed over where to place the knots in the flexible parametric models: precision was lower where the distribution of events in the smallest group differed substantially from the majority group but improved when the knots were forced to those in the minority covariate group.
The use of combined flexible parametric methods that fit age interaction as a spline term has shown promising results in the calculation of small populations. They can be used to model life expectancy by exact age, prediction of different covariate patterns under certain assumptions, and have greater statistical precision than traditional approaches. The implications of our findings to policymakers and researchers are discussed.

S32.2 Functional Limits of Agreement using a Mixed Effects Modelling Framework in Method Comparison Study

Kishor Das[1]

[1] *National University of Ireland, Galway, Ireland*

In any scientific study, the validity and reliability of a method of measurement is of crucial importance as the credibility of the findings depend upon it. Often a new method of measurement emerges that is more convenient compared to the existing method. A method comparison study is necessary to establish whether the agreement between the new method of measurement and the existing one is acceptable to warrant its use in general.
Statistical approaches for method comparison studies typically involve continuous (univariate) response variables. Little attention has been paid to the analysis of functional responses in method agreement studies where the response variable is functional in form e.g. comparing methods for continuous monitoring of blood pressure.
In method agreement studies involving a continuous univariate response, Bland and Altman proposed the use of 95% limits of agreement (LoA) to quantify the level of agreement between two methods of measurement and explore how individual measurements by the two methods vary within and between subjects. Due to the simplicity and usefulness of this approach it has received over 60,000 citations to date. This approach however is not directly applicable for method agreement studies involving functional responses. The complexity arises due to the functional nature of the response and when the functional responses are collected in hierarchical study designs.
In this presentation a novel approach using the mixed effects modelling framework is presented for the analysis of method comparison studies with functional responses allowing the generation of 95% Functional Limits of Agreement. The proposed method is very flexible and can accommodate hierarchical designs with replicates and covariates. Examples will be given including a comparison of marker-based and markerless motion capture systems in elite sports.

S32.3 Progression models for imaging data with Longitudinal Variational Auto Encoders

Benoît Sauty[1], Stanley Durrleman[2]

[1] *INRIA, Paris, France,* [2] *Inria Paris, Paris Brain Institute, France*

Disease progression models are crucial to understanding degenerative diseases. Mixed-effects models have been consistently used to model the evolution of clinical assessments or biomarkers extracted from medical images over time. They provide a statistical framework in which all patients are aligned on a common pathological timeline, and individual trajectories are parametrised as small variations (called random effects) around a reference trajectory (called fixed effects) that can be seen as the average scenario. This yields an interpretable description of the effects of time on each individual.
However, few progression models for entire medical images have been proposed. In this work we combine a dimension reduction approach, using a Convolutional Neural Network architecture called Variational Auto Encoders, with a statistical model, namely a linear mixed-effect model. The objective is to learn a low-dimensional latent representation of the high-dimensional images such that individual trajectories follow straight lines over time and are characterised by a few interpretable parameters. A longitudinal loss for the neural networks is proposed to ensure that the learned latent representations of images are compatible with the latent linear model, and a Monte Carlo estimator is devised to iteratively optimise the networks and the statistical model. This generative model allows to predict patients' future evolution at any timepoint, and to impute missing data, which is important in the context of disease modeling where visits are often missing or irregularly spaced.
We first validate this method on a synthetic data set to illustrate the disentanglement between time dependant changes and inter-subjects variability, as well as the predictive capabilities of the method. We then apply it to two different 3D imaging modalities, MRI and FDG-PET scans from the Alzheimer's Disease Neuroimaging Initiative (ADNI), to recover well documented patterns of structural and metabolic alterations of the brain in normal and pathological ageing. Using the raw images alleviates the need for heavy processing to extract imaging biomarkers, which does not only save time but also makes the approach independent of prior choice of biomarkers.

Parallel Sessions

S32.4 Modelling the effect of time varying covariates in time to event studies of twins with delayed entry

Annah Muli[1], Mar Rodríguez Girondo[2], Jeanne Houwing-Duistermaat[1]

[1] University of Leeds, United Kingdom, [2] Leiden University Medical Centre, Netherlands

The steady increase of life expectancy has led to increasingly aged populations in industrialized countries. Consequentially, it exists an increasing interest in studying health age-related outcomes in elderly populations and identify biomarkers associated with these outcomes. For example, fracture incidence is a public health problem in elder women. In this work, we aim to elucidate the association between the changes in bone mineral density (BMD) on fracture incidence in women older than 50 using data from the UK Twin Registry (TwinsUK). Age at first fracture is defined as the outcome of interest and BMD is a longitudinal biomarker measured multiple times during follow-up. Extra novel methodological challenges are the clustered nature of the twin data and dealing delayed entry (not all twins enter the study at the same age). We adopt a shared frailty approach to model the clustered nature of twins' time-to-event data. For singletons methods exist to estimate the parameter representing the effect of the time varying covariate on the outcome in data subject to delayed entry. However, methodology combining clustered time to event data, longitudinal biomarkers and delayed entry is lacking, even under the simplest assumption of the classical time varying covariates models.
We developed and compared two methods: first, we implemented an extension of the classical time varying covariates model to our setting with clustering and delayed entry. Second, we implemented a two-stage approach based on risk-set regression calibration. We will present the results of a comparative simulation study of both methods. Methods are compared assuming different levels of measurement error and different measurement patterns for the biomarker. Based on the simulations, the two stage approach performs well for variance of measurement error below 0.3 and overall performs better than the classical time varying covariates model. We will present the results of our application to the TwinsUK data (374 female dizygotic twin pairs). For BMD we have on average 3 measurements per individual. The estimate for the BMD effect is −3.515 based on the two stage approach and -3.117 based on the classic time varying covariates model. We will conclude by discussing potential advantages and disadvantages of the proposed methods and future research directions.

S32.5 Quantitative prediction error analysis to investigate performance under measurement heterogeneity

Kim Luijken[1,2], Jia Song[1], Rolf Groenwold[1]

[1] Leiden University Medical Center, Netherlands, [2] University Medical Center Utrecht, Netherlands

BACKGROUND: When a predictor variable is measured in similar ways at the derivation and validation setting of a prognostic prediction model, yet both differ from the intended use of the model in practice (i.e., 'predictor measurement heterogeneity'), performance of the model at implementation needs to be inferred. This study proposed an analysis to quantify the impact of anticipated predictor measurement heterogeneity.
METHODS: A simulation study was conducted to assess the impact of predictor measurement heterogeneity across validation and implementation setting in time-to-event outcome data. The use of the quantitative prediction error analysis was illustrated using an example of predicting the 6-year risk of developing type-2 diabetes with heterogeneity in measurement of the predictor body mass index.
RESULTS: In the simulation study, calibration-in-the-large of prediction models was poor and overall accuracy was reduced in all scenarios of predictor measurement heterogeneity. Model discrimination decreased with increasing random predictor measurement heterogeneity.
CONCLUSIONS: Heterogeneity of predictor measurements across settings of validation and implementation reduced predictive performance at implementation of prognostic models with a time-to-event outcome. When validating a prognostic model, the targeted clinical setting needs to be considered and analyses can be conducted to quantify the impact of anticipated predictor measurement heterogeneity on model performance at implementation.

Poster Presentations

ISCB43 NEWCASTLE,UK | 2022

Poster Presentations

P1

High dimensional data

«Great in, great out»: how to select a subpopulation from data with no response variable

Lubomir Stepanek[1], Filip Habarta[2], Ivana Mala[2], Lubos Marek[2],

[1] *First Faculty of Medicine, Charles University, Prague, Czech Republic,* [2] *Faculty of Informatics and Statistics, Prague University of Economics and Business, Czech Republic*

When having more initial data than needed, selecting a subpopulation from the given dataset is appropriate. However, a random subsampling could not always ensure that all categorical variables' levels are equally covered and all numerical variables are well balanced. When a response variable is included in the data, propensity scoring helps detect which subset of the original tends to keep initial data homogeneity and the variables' effect sizes harmonized. However, when a response variable is missing, the logistic model behind the propensity scoring cannot be built.
This study addresses the subpopulation selection with the missing response variable about to be collected later, using a COVID-19 dataset (N = 700) with 20 variables of interest, which was reduced to a subdataset (n = 400) by the following methods. First, numerical variables were categorized. Then, the quality of subpopulation selection was measured using a sum of squares of each variable's category frequency and averaged over all variables. Minimizing the metric reflects the demand for keeping all the variables' categories numerically balanced, i. e. of similar sizes. Several subset-selecting strategies were applied. Besides a single random subsampling, an exhaustive method selecting all possible combinations of n = 400 observations from initial N = 700 observations was performed, choosing the subsample and minimizing the metric. Similarly, a «forward» subselection reducing the original dataset observation by observation per each step by permanently lowering the metric was done. A repeated random subsampling enabled the model of a prior distribution of the metric and helped estimate its empirical minimum, determining one given subsample. Finally, k-means clustering (with a random number of clusters) of the original dataset's observations and choosing a subsample from each cluster, proportionally to its size, also lowered the metric compared to the random subsampling.
All four latter approaches showed better results than single random subsampling, considering the metric minimization. However, while the exhaustive sampling is very greedy and time-consuming, the forward one-by-one reducing the original dataset, picking up the subsample minimizing the metric, and subsampling of the clusters, are feasible for selecting a well-balanced subdataset for e. g. purposes of machine-learning training.

P2

Miscellaneous

A Bayesian approach for DBPCFC dose-response analysis of multiple allergens

Alma Revers[1], Michel Hof[1], Koos Zwinderman[1]
[1]*Amsterdam University Medical Center, Netherlands*

The double-blind placebo control food challenge (DBPCFC) is the gold standard test for food allergens. A subject with a suspected food allergy is given increasing doses of the allergen on the active challenge day and a placebo on the control day. After each dose, symptoms are recorded. Main question of interest is the difference between the two days with respect to the dose-response curves. As an example we analyzed DBPCFC-data of 398 patients on 9 different allergens.
Per allergen symptoms were categorized into 3 categories: no symptoms, only oral allergy syndrome or more severe symptoms. We estimated the dose-response curves per individual using random-effects models. As a subject may have symptoms on the control day, we adjusted the dose-response distributions based on the response of the control day.
Response-probabilities per allergen were modelled using proportional odds models with either the logit or the probit-link functions or with a multinomial logistic model without a proportional odds assumption. Each model had subject-specific intercepts and regression coefficient(s) to model the effect of the challenge-doses. For the placebo day the regression coefficients of doses were fixed to zero. Subject-specific coefficients were assumed to be drawn from normal distributions. We used a Bayesian approach using MCMC using a Gibbs sampler to estimate parameters assuming weakly informative priors. Furthermore, we averaged over the 3 models using pseudo-Bayesian model averaging as implemented in the R package 'LOO' to estimate model weights and obtain model-averaged estimates of the dose-response curves.
Additionally, we combined the data of the 9 allergens using a hierarchical model. We hypothesized that dose-response curves of similar allergens should be comparable and in particular we hypothesized that the regression coefficients of the challenge doses of similar allergens were sampled from a common distribution. This assumption will facilitate estimation of these dose-parameters in case of sparse data as was the case with our example data. Results of the hierarchical model were compared with results of per allergen analyses.

P4

Simulation studies

A comparison of different estimands of causal effects in competing risks setting using a simulation

Emama Amin[1]

[1] *Institute of Statistical Research and Training, University of Dhaka, Bangladesh*

In time-to-event data, a competing event is any event that makes it impossible for the event of interest to occur. Failure to account correctly for competing events can result in adverse consequences, including overestimation of the probability of the occurrence of the event and mis-estimation of the magnitude of relative effects of covariates on the incidence of the outcome. Young et al. (2020) recently developed a counterfactual framework to explicitly define two estimands of causal effects in competing risk settings. Contrast of the risk that would have been observed if all individuals had been assigned to treatment and we had somehow eliminated competing events is known as direct effect.
Contrast of the risk that would have been observed if all individuals had been assigned to treatment without elimination of competing events is known as total effect.
In this study we aim to compare the performance of the estimands for estimating causal effect when competing event exists (direct effects and total effects). A simulation study has been performed to compare the performance of the estimands in terms of sample size, rate of observed event of interest and regression parameters. Further, to present an application of these estimands to a data from a randomized trial designed to estimate the effect of estrogen therapy on prostate cancer mortality was used. The results of simulation study concludes that, in terms of BIAS, one should choose to interpret total effect (Risk difference) and if the rate of observed event of interest is high, direct effects (both Risk ratio and Risk difference) can be used. In terms of efficiency, when interpreting total effect, risk difference should be used over risk ratio. Direct effects (both RR and RD) are more efficient than total effects (RR and RD) when rate of observed event of interest is higher.

P6

Simulation studies

A comparison of methods for estimating dichotomous treatment effects: a simulation study

Jacqueline Thompson[1], Samuel Watson[1], Lee Middleton[1], Karla Hemming[1]

[1] *University of Birmingham, United Kingdom*

BACKGROUND: A common approach for estimating covariate-adjusted binary treatment effects comparing a treatment and control group with dichotomous outcomes is the odds ratio (estimated via logistic regression) because of its stability and robustness to model mis-specification. However, other measures of treatment effect, including - relative risk and risk difference - are arguably easier to interpret and better suited to specific designs such as non-inferiority studies. However, there are no comprehensive comparisons of different candidate methods to estimate adjusted measures of relative risk and risk difference.
OBJECTIVES: A simulation study is used to evaluate the performance of a number of candidate methods used to estimate relative risks. We consider log-binomial, generalised linear models (GLM) with iteratively weighted least-squares (IWLS) and non-robust standard errors (SE); log-binomial GLM with convex optimisation and robust SEs; modified Poisson GLM IWLS and non-robust SEs; log-binomial generalised estimation equations (GEE) and robust SEs; marginal standardisation and delta method SEs; and marginal standardisation and permutation test SEs.
Performance measures (coverage, bias and convergence) are evaluated across scenarios covering a range of sample sizes, event rates, covariate prognostic strength, and model mis-specification. Datasets are simulated from a randomised controlled trial, replicated 10000 times for each scenario across all possible combinations of sample sizes (200, 1000, and 5000), outcomes (10%, 50% and 80%), and covariates (ranging from -0.05 to 0.7) representing weak, moderate or strong relationships. Null and alternative treatment effects (ranging from 0, -0.5, 1; on the log-scale) assess coverage and power, respectively.
RESULTS: Findings will be presented summarising estimation algorithm convergence rates, absolute mean bias, mean square error (MSE), coverage of 95% confidence intervals, and power.
CONCLUSION: There are several methods for estimating unadjusted and adjusted relative risks. However, it is unclear which method is the most robust or powerful when adjusting for covariates. Marginal standardisation and convex optimisation may perform better than GLM IWLS log-binomial. GEE estimations may be biased for outcome conditional outcome distributions.

Poster Presentations

P7

Missing data

A comparison of strategies for selecting auxiliary variables for multiple imputation

Rheanna Mainzer[1], Cattram Nguyen[1], John Carlin[2], Margarita Moreno-Betancur[2], Ian White[3], Katherine Lee[1]

[1] Murdoch Children's Research Institute, Australia, [2] University of Melbourne and Murdoch Children's Research Institute, Australia, [3] MRC Clinical Trials Unit at UCL, United Kingdom

Multiple imputation (MI) is a popular method for handling missing data. Auxiliary variables can be added to the imputation model(s) to improve MI estimates. However, the choice of which auxiliary variables to include in the imputation model is not always straightforward. Including too few may lead to important information being discarded, but including too many can cause problems with convergence of the estimation procedures for imputation models. Several data-driven auxiliary variable selection strategies have been proposed. We used a simulation study and a case study to provide a comprehensive comparison of the performance of eight auxiliary variable selection strategies, with the aim of providing practical advice to users of MI. A complete case analysis and an MI analysis with all auxiliary variables included in the imputation model (the full model) were also performed for comparison. Our simulation study results suggest that the full model outperforms all auxiliary variable selection strategies, providing further support for adopting an inclusive auxiliary variable strategy where possible. Auxiliary variable selection using the Least Absolute Selection and Shrinkage Operator (LASSO) was the best performing auxiliary variable selection strategy overall and is a promising alternative when the full model fails. All MI analysis strategies that we were able to apply to the case study led to similar estimates.

P9

Miscellaneous

A Flexible Copula Model for Bivariate Survival Data with Dependent Censoring

Reuben Adatorwovor[1], Yinghao Pan[2]

[1] University of Kentucky, United States, [2] University of North Carolina at Charlotte, United States

Censoring in time-to-event survival analysis is very popular and a useful way of incorporating available data in analysis of right-censored survival data. The existing literature mostly assumes that the exact event-time and the censoring time are independent and non-informative. An assumption that may be unsubstantiated and generally introduce biases in the estimation procedure. Popular methods like the Inverse Probability Weighting (IPW) and other variants attempt to correct for the biases associated with the independent censoring assumption in a univariate setting. We proposed a copula-based dependent censoring methodology for bivariate time-to-event data, which models the event and censoring times through a copula-based Cox Proportional Hazard model formulation. Our estimator possesses strong consistency and desirable asymptotic properties under regularity conditions. We provide results under extensive simulation and applied our method to the Danish twin's registry dataset.

P10

Miscellaneous

A lean additive frailty model: with an application to clustering of melanoma in Norwegian families

Mari Brathovde[1], Tron A. Moger[2], Marit B. Veierød[2], Tom Grotmol[3], Odd O. Aalen[2], Morten Valberg[1]

[1] Oslo University Hospital, Norway, [2] University of Oslo, Norway, [3] Cancer Registry of Norway

Large scale health registries enable detailed studies of clustering of cancers in families. Frailty models provide a framework for conducting such studies at a much higher level of detail than conventional analyses. However, the complexity of these models grows with both family size and number of cancer diagnoses in each family. Consequently, these models have largely been used in settings where cluster sizes are small, e.g. for twin pairs, or by considering only a few first-born children in a family. This poses a challenge for fully utilizing the detailed data available in the registries. We present a modification of the additive genetic gamma frailty model, which alleviates some of these problems by using a leaner additive decomposition of the frailty.

By using a first order approximation of the genetic structure in nuclear families of parents and children, we obtain a model that reduces the complexity with respect to family size. This allows us to analyze a far greater class of datasets. An additional major benefit of the lean model is the significant speed up in model fitting. This enables fitting even more complex models and makes model fitting on a desktop computer feasible, without the need of a high performance cluster (HPC).

Using the lean model, we were able to analyze a complete population-wide data set on melanoma in all 2,391,125 Norwegian families registered in 1960-2016. We found a substantial clustering of melanoma in Norwegian families, and a large heterogeneity in melanoma risk across the population. We estimated a considerable frailty variance of 2.25, which is reflected in familial relative risks, where there is a 2–2.4-fold and 4-fold increase in one's risk if one or two siblings, respectively, are affected by melanoma. We found a large inequality in frailty in the population, where 46% of the frailty could be attributed to the 10% of the population at highest unobserved risk.

In conclusion, additive frailty models can be used to study relatively large clusters. Furthermore, there is a substantial clustering of melanoma in Norwegian families, and a large heterogeneity in melanoma risk across the population.

P11

Efficient clinical trial designs

A multi-arm multi-stage platform design that allows pre-planned addition of arms while still control

Peter Greenstreet[1], Thomas Jaki[2], Alun Bedding[3], Chris Harbron[4], Pavel Mozgunov[2]

[1] Lancaster University, United Kingdom, [2] MRC Biostatistics Unit, University of Cambridge, United Kingdom, [3] Roche, [4] F. Hoffmann–La Roche

There is growing interest in platform trials that allow for adding of new treatment arms as the trial progresses as well as being able to stop treatments part way through the trial for either futility or for superiority. An important feature of many platform trials is the need to guarantee that error rates are controlled. This talk presents a multi-stage design that allows additional arms to be added in a platform trial in a pre-planned fashion, while still controlling the family wise error rate and achieving a desired level of power. A motivating trial is discussed which focuses on two key settings, with the first being a set number of stages per active treatment arm and the second being a set total number of stages, with later treatments getting fewer stages. These two settings are compared to running multiple separate trials and running a traditional multi-arm multi-stage design. Through this example we show that the proposed method results in a smaller sample size while still controlling the errors compared to running multiple separate trials. We also discuss how this approach can reduce the time until a treatment can enter the market compared to running a traditional multi-arm multi-stage design.

Poster Presentations

P13

Machine learning methods for health

A novel scoring system for diagnosis of Tuberculous Meningitis: a Bayesian Latent Class Analysis

Trinh Dong[1], Ronald Geskus[1]

[1] *Oxford University Clinical Research Unit, Vietnam*

BACKGROUND: Tuberculous meningitis (TBM) is the most lethal form of tuberculosis. There is no gold standard. There is a uniform case definition, which is based on expertise and comes with high uncertainty. We developed a model using latent class analysis (LCA) that provides a statistical scoring system estimating individual TBM risk amongst adults with suspected brain infection, taking into account the uncertainty of TBM diagnosis.
METHODS: The model contains two components: i) diagnostic test model relating latent TBM status to three tests as manifest variables, each providing different ways to detect bacteria. The local independence between manifest variables is taken into account via latent bacillary burden. Three different linear mixed-effect models are considered; ii) a diagnostic prevalence model relating latent TBM status to known risk factors. A Bayesian approach was used to incorporate prior knowledge and reduce the degrees of freedom. Missing data was dealt with depending on anticipated missing mechanisms. All considered models were validated using 5 repeated 20-fold cross validations. Hospital diagnosis made by physicians at the end of follow-up allowed us to perform pseudo validation of the prevalence model.
RESULTS: Our selected model showed good AUC of 92.4%, 92.9%, and 96.1% and good calibration (intercepts = 0.05, 0.1, 0.17 and slopes = 1.1, 1.2, 1.3) for the three tests. The prevalence model had an AUC of 93.9%, and a calibration intercept and slope of -0.01 and 0.97.
CONCLUSION: Compared to the uniform case definition, our model provides a well-calibrated scoring system with good correspondence to current clinical judgement. Our model also gives an insight into the mechanistic pathway from TBM to test results via disease bacillary burden at diagnosis.

P14

Communicating statistical methods

A parametric multiplicative rate model for recurrent gap time data with shared frailty

Ivo Sousa-Ferreira[1], Ana Maria Abreu[2], Cristina Rocha[1]

[1] *CEAUL and Faculdade de Ciências, Universidade de Lisboa, Portugal,* [2] *Faculdade de Ciências Exatas e da Engenharia, Universidade da Madeira and CIMA, Portugal*

In longitudinal studies, it is common for subjects to experience several episodes of a certain event of interest during the observation period. Such outcomes have been termed recurrent events. In many settings, researchers have more interest in gap times than in the time-to-events. For instance, in medical studies it is often intended to model the gap times when subjects can recover after each event, such as in asthma attacks, epileptic seizures, among others. In this context, Zhao and Zhou [1] proposed a semi-parametric rate model that is derived from a non-homogeneous Poisson process. However, parametric modelling is more convenient when the research interest is also to study how the recurrence rate evolves over time. Recently, Sousa-Ferreira et al. [2] proposed the extended Chen-Poisson (ECP) distribution in order to specify the baseline rate function. The proposed model turns out to be quite flexible, since it enables complex shapes for the rate function. Although this model allows an adequate modelling of the gap times distribution, it is not able to deal with the unobserved heterogeneity, which is a source of within-subject correlation, given that subjects can be more (or less) prone to experience their subsequent events due to unknown or unmeasurable risk factors. Thus, in this work we extend the ECP rate model by incorporating a frailty term that is shared across the gap times of each subject. The frailty variable is introduced multiplicatively on the rate function, with the purpose of representing the association between the subject's gap times. Moreover, two alternative distributions for the frailty are investigated, namely the gamma and inverse Gaussian distributions. The maximum likelihood method is applied for parameters estimation in the presence of right-censoring. An application to clinical data is considered to illustrate the potential of the new shared frailty model.

REFERENCES:
[1] Zhao, X., Zhou, X. Modeling gap times between recurrent events by marginal rate function. Computational Statistics & Data Analysis, 56(2), 370-383, 2012.
[2] Sousa-Ferreira, I., Rocha, C., Abreu, A.M. The extended Chen-Poisson marginal rate model for recurrent gap time data. In Recent Developments in Statistics and Data Science, Bispo, R., Rodrigues, L., Alpizar-Jara, R., de Carvalho, M. (eds.), Springer Proceedings in Mathematics & Statistics. Accepted for publication.

P15

Machine learning methods for health, High dimensional data

A Prediction of diabetic retinopathy with machine learning and time-varying covariates

Panu Looareesuwan[1], Suparee Boonmanunt[1], Thitiya Lukkunaprasit[2], Sukanya Siriyotha[1], Ratchainant Thammasudjarit[1], Ammarin Thakkinstian[1], Thunyarat Anothaisintawee[1]

[1] *Mahidol University, Thailand,* [2] *Rangsit University, Thailand*

Type 2 diabetes (T2D) is one of the major non-communicable diseases with an estimated global prevalence in adults was about 10%. It was predicted to be about 12% by 2045 which is about 600 million people globally affected. Diabetic retinopathy (DR) is very common among type 2 diabetes (T2D) where the global prevalence is estimated to be about 22.27%. It is reported that about 60% of individual will develop DR after 20 years of diagnosis of T2D. In Thailand, the prevalence of diabetic retinopathy, non-proliferative DR (NPDR), mild NPDR, moderate-severe NPDR, severe NPDR and proliferative DR (PDR) was 24%, 9.4%, 10.5%, 1.3% and 2.8%, respectively. There are many well-developed traditional statistical models to predict the progression of DR. However, most of the traditional statistical modeling requires strong assumptions of constant hazard and interaction between features needs to be assessed. Also, many of the features that contribute to the disease progression vary over time. Machine learning (ML) models are capable of handling non-linear features and do not require proportional hazard assumption. The dataset used in this study is Real-world data (RWD) of clinical practice between 2010-2019 from Ramathibodi, a university hospital, Thailand with a sample size of 48,622 individuals. The objective of this study is to compare the prediction performance of the traditional statistical Cox-Proportional Hazard (CPH) model with ensemble ML models, left-truncated right-censoring relative risk forest (LTRCRRF) and left-truncated right-censoring conditional inference forest (LTRCIF). These models are applied with time-dependent features. The dataset was split into train, validate and test datasets by subjects. The ML models were trained on the training set and the best hyperparameters were selected with a validation set. The performance shows that CPH model has superior Harrell's C-index of 0.72, 0.73, 0.69 in a train, validate and test sets. In contrast, LTRCRRF and LTRCIF C-index were 0.69, 0.60, 0.59 and 0.57, 0.54, 0.52, respectively. This shows that the performance of the traditional statistical models on the tabular data is still superior and explainable than ML models.

P17

Efficient clinical trial designs, Personalized medicine

A review of early phase dose-finding clinical trials with incomplete follow-up for toxicity consider

Zhulin Yin[1], Christina Yap[1], Adrian Mander[2]

[1] *Institute of Cancer Research, United Kingdom,* [2] *Cardiff University, United Kingdom*

PURPOSE: Incomplete follow up for toxicity caused by delayed toxicity or rapid enrolment of participants is a challenge for early phase dose-finding trials. We conducted a comprehensive review of the early phase clinical trials that used designs accounting for partial and complete toxicity information, with the aim of understanding how these designs have been implemented in practice and investigating the reporting quality of these trials.
METHODS: We identified 108 trial papers published between January 2004 and December 2020 that used the time-to-event continuous reassessment method (TITE-CRM), TITE-CRM with cycle information, the Rolling six design (Rolling 6), the time-to-event cumulative cohort design (TITE-CCD) and the rapid enrolment design (RED). Clinical settings, design parameters, assumptions on the toxicity profile, interim analysis and practical considerations were extracted from the published papers and protocols (when available).
RESULTS: Most trials used TITE-CRM (47, 43.5%), Rolling 6 (55, 50.9%), TITE-CRM with cycle information (3, 2.8%), TITE-CCD (2, 1.9%) and RED (1, 0.9%). The number of trials implementing these designs has increased over time and all of them are oncology trials. Trials using TITE-CRM tended to have a longer DLT assessment window compared to the Rolling 6 design (32(64.0%) beyond the first cycle of treatment (C1) vs 50 (90.9%) within C1). More paediatric trials are conducted with Rolling 6. All trials using the Rolling 6 design provided sufficient information to enable replication of the interim and final analyses. However, the reporting quality of design parameters for TITE-CRM is poor. Only four trials implementing TITE-CRM provided sufficient details to enable a fully replicable final analysis. Of those trials, only one reported all the required information to enable replication of both the interim and final analyses.
CONCLUSION: We found deficiencies in the reporting of essential details in the implementation of TITE-CRM. The failure to report design parameters and trial results fully and accurately, may do needless harm to the credibility of more complex trial designs and clinical research. A considerable improvement is needed. We provided recommendations to improve transparency, reproducibility and to enable accurate interpretation of results for such designs.

Poster Presentations

P18

Simulation studies, Meta-analysis

A simulation study to compare methods for estimating the sample mean and standard deviation

Ram Bajpai[1], Joie Ensor[1], Dean Lake[1]

[1] *Keele University, Newcastle, United Kingdom*

BACKGROUND: In a meta-analysis of continuous outcomes, mean and standard deviation (SD) are commonly pooled from different studies to generate a combined effect estimate. However, studies may report median and related measures of dispersion instead of sample mean and SD for a given outcome where the outcome follows a skewed distribution. Therefore, in order to combine individual study results in a consistent format, it is necessary to estimate the sample mean and SD for such studies.
OBJECTIVE: To compare the relative performance of methods for estimating the sample mean and SD from the sample size, median, range and/or interquartile range for different scenarios.
METHODS: In this methodological review, we investigate existing methods for estimating sample mean and SD using other available summary statistics. A number of medical statistics methodology journals such as BMC Medical Research Methodology, Statistical Methods in Medical RESEARCH, Research Synthesis Methods, and Statistics in Medicine were searched to identify proposed methods for estimating sample mean and SD. We also searched Google Scholar for grey literature to locate potentially relevant articles. In a given study, summary data is most commonly reported as either; S1 (a, m, b; n), S2 (a, q1, m, q3, b; n), or S3 (q1, m, q3; n) where a = minimum, m = median, b = maximum, q1 = first quartile, q3 = third quartile, and n = sample size. A simulation study was conducted to assess the performance of each existing method in terms of relative error and root mean squared error (RMSE). We generated simulated data from various distributions including assuming normality and various skewed distributions such as lognormal, exponential, Weibull, and beta distribution for different sample sizes. We also compare the performance of the different methods in different scenarios using a real dataset.
CONCLUSION: Initial results using the simulated data from the Weibull distribution showed that no single method fits to all scenarios of estimating mean and SD. Therefore, it is essential to know and inform meta-analysts about which proposed method works better in commonly reported scenarios and under particular distributional assumptions to avoid any poor estimation of the sample mean and SD.

P20

Personalized medicine, Missing data

A workflow to perform MAIC with multiple imputation applied to the ESME database and aggregated data

Cyril Esnault[1], Vanessa Barbet[2], Thomas Filleron[3], Gaëlle Chenuc[4], Maurice Pérol[5], Didier Debieuvre[6], Nicolas Girard[7], Xavier Quantin[8], Katia Thokagevistk[1], Gaëtane Simon[9], Louise Baschet[2]

[1] *Medical Data Center & Personalized Medicine, Roche, Boulogne-Billancourt, France,* [2] *Horiana, Bordeaux, France,* [3] *Institut Claudius Régaud IUTC-O, Toulouse, France,* [4] *Biometry department, IQVIA, France,* [5] *Medical Oncology Department, Centre Léon Bérard, Lyon, France,* [6] *Chest disease department, GHRMSA, Mulhouse, France,* [7] *Medical oncology department, Institut du Thorax Curie-Montsouris, Paris, France,* [8] *ICM and IRCM, INSERM U1194, University of Montpellier, Montpellier, France,* [9] *Health Data and Partnership Department, Unicancer, Paris, France*

BACKGROUND: Matching-adjusted indirect comparison (MAIC) provides a method for comparing absolute treatment effects with limited availability of individual patient data, by matching the baseline characteristics of the trial populations. In real world studies, multiple imputations (MI) are often used to replace missing values with a set of plausible values that represent the uncertainty about the right value to impute. In the context of small sample sizes, both processes of MAIC and MI combined can lead to methodological issues and convergence problems. The aim was to develop a workflow to anticipate the different steps until the obtention of an acceptable model. This approach was applied to ROS1-positive metastatic non-small cell lung cancer patients, with the comparison of entrectinib aggregated clinical trials data and the French national ESME-AMLC cohort.
METHODS: This study was conducted in 3 phases (go/no go): 1) feasibility assessment, 2) weights estimation and 3) outcome and inference analyses. This abstract focuses on phase 2. The key prognostic factors were age, gender, ECOG (~45% missing), tumor histology, smoking status (~6% missing), brain metastases and the number of previous treatment lines. The proposed workflow is a 4-steps sequential implementation for calculating for each imputed dataset the weights using logistic regression (moments' method):
1- Initial models with all predictors;
2- If no model convergence for all datasets, simplification by performing predefined steps: a) removing variables with excessive major category in both groups, b) refactoring variables;
3- Weights truncation above 0.01;
4- Backward procedure on imbalanced variables until reaching all means Standardized Mean Differences by variable < 0.15, with a) refactoring, b) removal if it allows balance on other variables.
RESULTS: Five populations in the ESME database were considered, 2 of them reflecting French HTA recognized comparators, the other 3 the clinical practice (between 19 and 70 patients). The workflow allowed selecting for each population, the model that converged for all imputed datasets, with balanced baseline characteristics. Effective sample sizes after weighting ranged between 10 (-47% versus initial sample size) and 60 (-14%).
CONCLUSION: This approach, developed in a specific case in oncology, can be extrapolated for other studies.

P22

Machine learning methods for health, Prediction models

An approach for constructing decision tree using density estimates

Asanao Shimokawa[1], Etsuo Miyaoka[1]

[1] *Tokyo University of Science, Japan*

In the field of medical research, the prediction of interesting outcome based on patient information is an important problem. Tree-structured models like decision tree or survival tree are widely used as one choice to achieve this problem, and the classification and regression tree (CART) algorithm is a popular method for constructing these models. The CART consists of splitting, pruning, and selection steps. These steps allow the automatic construction of optimal combination of covariates and the set of rules to construct the subsets of covariates space. A key setting in the algorithm is how to measure the reduction of impurity in a node due to splitting of covariate space and creating two child nodes.
The reduction of impurity in a node due to splitting is estimated by using the estimates of the probability that any case falls into a child node and predetermined criteria of impurity within a node like mean squared error or gini index. Although there are extensive researches on the effect of differences between criteria of impurity within a node, little research has been done on the effects of differences between estimation methods of the probability that any case falls into a child node and simple resubstitution estimate is widely used. Therefore, our goal is to investigate the effect of differences between estimation methods of the probability that any case falls into a child node. Estimating about the probability that any case falls into a child node is equivalent to estimating about the covariate distributions in a node. Therefore, we propose a new splitting and pruning method using the kernel density estimate of covariates in a node and evaluate the performance of this through simulation studies. From the simulation studies, the trees obtained by proposed method had higher performance than the trees obtained by conventional approach in several situations.

P23

Efficient clinical trial designs

An on-going two-stage adaptive randomized controlled trial of Andrographis and Boesenbergia extract

Pawin Numthavaj[1], Amarit Tansawet[1], Atiporn Ingsathit[1], Sasisopin Kiertiburanakul[1], Suradej Hongeng[1], Phisit Kemawoot[1], Ammarin Thakkinstian[1]

[1] *Faculty of Medicine Ramathibodi Hospital, Thailand*

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) or COVID-19 has infected approximately 190 million people around the world. Even though most of the COVID-19 infections are asymptomatic, 5-10% of the cases developed severe symptom such as pneumonia. Andrographis paniculata, well-known herbal medicine widely used in Thailand, contains active ingredient called andrographolide. This herb has been commonly used for relief of the upper respiratory tract symptoms such as fever and sore throat and common cold. The efficacy of andrographolide is possibly explained by its anti-inflammatory effect. Another potential herbal medicine is Boesenbergia rotunda extract, which has panduratin A as an active ingredient. Currently both herbal medicines are being used by local Thai people without proper evidence to support the use. Therefore, we aimed to conduct an RCT to evaluate the benefit of both treatments. The ongoing study is an open-label, three-armed, parallel, two-stage, adaptive RCT of Andographis extract, Boesenbergia extract, and standard treatment with mild to no COVID-19 symptoms admitted in hospitel in Bangkok, Thailand. Stage 1 is expected to recruit patient about 1704 patients, or 568 per group. Patients are receiving either Andrographis extract, Bosensurgia extrat, or standard supportive care, and then followed up for 4 weeks. Stage 2 will be conducted after the result has been analyzed by an independent biostatistician who is not involved directly with the study, with the approval of the data and safety monitoring board. Expected total sample size is around 3060. Primary outcome of the study is the discharge status, i.e. discharge to home or hospitalization. Additionally WHO clinical progression scale is also assessed as the secondary outcome at 4 weeks, as well as side effects from the treatment. Currently the study has been approved by the ethics committee, and recruitment has been started with 24 patients completed the follow-up with no severe adverse event. This research is supported by Health Systems Research Institute of Thailand, Ministry of Health.

Poster Presentations

P25

High dimensional data

Analysis of concentration-time-response data in cell-based compound profiling

Michal Kahle[1]

[1] *IKEM, Prague, Czech Republic*

Compound profiling is performed during drug discovery to identify interactions of candidate drugs with diverse molecular targets. This may in turn help to elucidate the mechanism of action and predict toxicity, side effects and safety issues of the molecules of interest. One of the methods of compound profiling is the monitoring of changes in various aspects of morphology and metabolism of laboratory cell lines upon exposure to the chemicals and clustering the responses based on similarity.
We have used a system measuring the impedance of cells growing on top of arrays of tiny golden electrodes (xCELLigence RTCA, Agilent) to measure response profiles of 6 human cancer cell lines to 16 concentrations of 122 chemical compounds measured in 44 time points over 72 hours. This gives 4224 data points for each compound. Here we compare the quality of hierarchical clustering of the full data, data with reduced dimensionality by PCA and UMAP, and our novel concentration-time-response surface (CTRS) model.
CTRP uses traditional logistic dose-response curve (DRC) in the dimension of concentrations and Chebyshev polynomials of the first kind to model the evolution of DRC parameters in time. The model was fitted by slightly modified ridge regression. To achieve potency invariance of the model we have removed the parameters for the first (constant) Chebyshev polynomial. Clustering CTRS models gave consistently the best quality scores (AMI, ARI) across a wide range of cluster counts (k). In conclusion, we describe a novel way of modelling concentration-time-response data which can be useful in compound profiling campaigns to enrich the profiles in the time and concentration dimensions.

P26

Miscellaneous

Application of Item Response Theory to Examination of the Psychometric Properties of the MNA Scale

Michalina Luśtyk[1], Katarzyna Zawisza[1]

[1] *Jagiellonian University - Medical College, Krakow, Poland*

The quality of results obtained in research depends largely on the quality of the research tools used. Hence, it is extremely important to assess the validity and reliability of the measurement scales used. At the same time, there is a continuous development of statistical techniques applied in this area, especially techniques developed within the framework of the Item Response Theory (IRT). Therefore, the aim of the present study will be to apply the Generalized Partial Credit Model (GPCM) to assess the validity and reliability of the Mini Nutritional Assessment (MNA). The choice of model was motivated by the existence of ordinal categories of responses categories and their different number in individual test items.
The analysis will be based on the results of the cross-sectional study «Neglect and self-neglect older people - challenges for formal and informal caregivers and for professionals medical and social professionals of the health care and social assistance system» conducted in 2017 on a representative sample of older people (65+) from the area of Lesser Poland.
The slope parameters will be presented, which determine the coefficient of discriminatory power of a given item and the difficulty parameters, which, calculated for each item of the test, determine the dependence of a correct answer to the intensity of the trait under test. Information curves for the test items will be used to illustrate reliability.
Analyses will be performed using SAS 9.4 software.

P29

Ageing

Application of the Item Response Theory in the development and validation process of a self-neglect

Katarzyna Zawisza[1]

[1] *Jagiellonian University – Medical College, Krakow, Poland*

There is increasing interest in using Item Response Theory (IRT) in the scale development and validation process, due to its advantages over the Classical Test Theory (CTT) method. In the CTT the final score is usually obtained based on the sum of points which have been allocated to specific responses across the items and, consequently, both the distance between answer categories and the importance of each item are the same, which does not reflect reality. The IRT method makes it possible to take into account the amount of information and the intensity required to achieve a certain level of trait (self-neglect in our scales) in the final score calculation. Thus, the study aimed to use the IRT to develop and validate a tool to evaluate the level of self-neglect in community dwelling older people in Poland, based on both subjective and objective assessments.
The cross-sectional study of elder neglect and self-neglect was conducted in Lesser Poland in 2017. It included 2,894 face-to-face interviews with randomly selected community-dwelling individuals. The Generalized Partial Credit Model (GPCM) from the IRT was applied with R software using the ltm package. The GPCM model was chosen because it can handle polytomous ordinal manifest items with a different number of answers.
There were three scales developed: the Self-Reported Self Neglect Scale (SRSNS) to measure the subjective assessment of self-neglect, the Objective Assessment of the Level of Self-Neglect - Physical Appearance (OALSN-PA) scale, concerning physical health risks based on the appearance of an individual, and, the Objective Assessment of the Level of Self-Neglect - Standards of Living Arrangements (OALSN-SLA) scale, which assesses the physical and personal living conditions. The final scores of all the developed scales were computed based on the GPCM model, which allows to establish individuals' latent trait level based on their response patterns. The GPCM model enables to determine the probability of latent trait as a function of observed patterns of responses and item parameters (thresholds and slopes, which point out the item's features). The latent trait value that has the highest likelihood becomes the latent trait estimate.
The presented research provides an example of application of the GPCM model and gives the evidence for good quality of the developed scales to measure self-neg.

P30

Miscellaneous

Assessing the effect of a cluster-level intervention without randomisation using baseline data

Andrew Copas[1], Tim Morris[1]

[1] *MRC Clinical Trials Unit at University College London, United Kingdom*

In an ongoing study in Jharkand, India, a community intervention to reduce malnutrition has been implemented in six intervention districts and nine other districts were selected to serve as controls. A baseline survey was carried out and a cross-sectional 'endline' survey will follow in 2025, both in infants aged 0-36 months. The primary outcome is the weight-for-height Z-score (continuous).
In cluster randomised trials, balance at baseline is expected. Whilst there are different analysis methods, at least for a continuous outcome these typically target the same estimand. However in controlled interventional studies such as our example, with baseline data but without randomisation, methods can differ more fundamentally. When using the commonly applied 'difference in differences' method, the estimand is the difference between intervention and control arms in the average change over time. We argue a more relevant estimand is the (hypothetical) difference in average outcome at endline had there been balance between intervention and control arms at baseline.
We propose a novel alternative method based on a fitting a mixed model to the data from both baseline and endline and using it to predict, as intervention effect, the difference in the expected outcome at endline between clusters in intervention and control arms but with the same underlying outcome mean at baseline. In this model each cluster has a random effect for baseline and one for follow-up, which are correlated, and we include fixed effects for time, intervention arm, and their interaction. We explain how, for certain model structures, the estimand is a simple function of the model parameters. Consequently estimation of the effect, and calculating a confidence interval, are straightforward after fitting the model. This result extends previously published methodology for estimating how patient characteristics are associated with change over time in a biomarker accounting for difference in baseline values.
We contrast methods and explain their application to our example. We also consider the potential of weighting methods in this context. We compare method performance. The novel method is efficient and simple to apply, though checking model fit is required.

Poster Presentations

P31

(Semi-)competing risks and causal inference, Prediction models

Assessment of prognostic model performance: A competing risks cause-specific hazards approach

Sarwar Mozumder[1], Paul Lambert[1,2], Mark Rutherford[1], Sarah Booth[1]

[1] *University of Leicester, United Kingdom,* [2] *Karolinska Institutet, Solna, Sweden*

BACKGROUND: In order to ensure appropriate validation of a prognostic model, calibration is performed on absolute risks to assess the agreement between the predicted risks from the model and the observed risk. In the presence of competing risks, correct specification of more than one model may be required in order to ensure well-calibrated predicted risks for one cause. Furthermore, interest may be in the predicted risks of the cause of interest, competing events and all-causes. Therefore, calibration must be ensured on various measures simultaneously. Finally, calibration can be assessed via internal, or external validation, both of which will be explored.
METHODS: In the presence of competing risks, we argue that prediction models should be developed on the cause-specific hazards scale where miscalibration can be conveniently assessed using net probabilities of death. Calibration plots are presented on cause-specific cumulative incidence functions and net probabilities of death alongside performance measures such as the Brier Score and Index of Prediction Accuracy. Flexible parametric models are further proposed for easy estimation of the baseline cause-specific hazard for prediction of individual risk.
RESULTS: Simulated scenarios illustrate that, miscalibration due to a mis-specified covariate functional form in the prediction model, or changes in the baseline cause-specific hazard in external validation data are better identified on the net probabilities of death. A miscalibrated model on one cause, could lead to poor calibration on predicted risks for each cause of interest, including the all-cause absolute risk. This is because prediction of a single cause-specific absolute risk is impacted by effects of variables on the cause of interest and competing events.
CONCLUSIONS: When competing risks are present, accurate predictions on both all-cause and cause-specific absolute risk must be ensured. However, the effect of cause-specific predictors will affect each cause of absolute risk simultaneously. Therefore, miscalibration can more easily be identified using net probabilities of death using calibration plots and appropriate performance measures such as the Index of Prediction Accuracy.\

P32

Missing data

Asthma classification; framework of multiple imputation in cluster analysis

Halehsadat Nekoee Zahraei[1], Renaud Louis[1], Anne-Françoise Donneau[1]

[1] *University of Liege, Belgium*

Asthma is widely accepted as a complex and heterogeneous disorder with different underlying clinical phenotypes. These phenotypes may determine using several clinical approaches, however, the common statistical approach in this field is cluster analysis. In recent years, clustering in asthmatic patients has been extensively studied, however, this study recognized evidence of heterogeneity in one of the well-known asthma phenotypes, eosinophils, and non-eosinophils.
In spite of similar studies, the present study included comprehensive multi-dimensional clinical variables and required the use of dimension reduction techniques, such as principal component analysis. In order to perform dimension reduction, a complete dataset without any missing values must be available. This problem was addressed using multiple imputation. On the new components of each imputed dataset, a clustering procedure was applied. The final cluster was constructed by combining the ensemble of partitions using the proposed method, mixture multivariate multinomial model (4M). The proposed framework for clustering multi-dimensional variables with missing values could be extended to similar incomplete datasets. A variety of scenarios on simulated datasets inspired by real data with known clustering were analyzed to evaluate the framework's validity on incomplete datasets. The proposed method addressed some limitations in statistical literature to find high discriminating clusters.
The proposed framework, based on multiple imputation, variable reduction, and our proposed combination method for clustering, was then applied to both eosinophils, and non-eosinophils datasets from the Pneumology Department of the University Hospital of Liege, which aimed to identify clinical sub-clusters in both phenotypes. After extensive clinical characterization of the patients, the clustering on eosinophils and non-eosinophils phenotypes yielded two clusters that mainly differentiate by demographics, atopic status, level of asthma control, and functional and inflammatory features.

P33

Efficient clinical trial designs, Simulation studies

Bayesian decision making in bioequivalence trials with pilot data

Duo Lyu[1], Xinyue Zhang[2], Michael Grayling[2], Haiyan Zheng[3]

[1] *The First Affiliated Hospital, Zhejiang University School of Medicine, China,* [2] *Newcastle University, Newcastle upon Tyne, United Kingdom,* [3] *University of Cambridge, United Kingdom*

In the verification of bioequivalence between a generic drug and its branded counterpart, a pilot trial with only a few patients would often be conducted before a pivotal bioequivalence trial. The accrued pilot data are used to assess the feasibility and to estimate the range of possible effects. Due to limited sample sizes, it is not recommended to decide if a pivotal should occur based on significance tests at the conventional level of 5%. Whilst some authors suggest adjusting the significance level in such analyses, the Bayesian framework provides an alternative approach to inform a Go/No-go decision. Moreover, it brings about the possibility of incorporating pilot data in a prior for parameters that underpin the pivotal trial.
Consider two-sequence, two-period crossover designs that compare the test (T) and reference (R) treatments. We propose a robust Bayesian hierarchical model to accommodate the respective linear mixed-effects models fitted to the pilot and pivotal trials. The study-specific treatment effects (δi = μTi - μRi, i = 1, 2) can further be modelled with a random-effects distribution. Robust prediction is attained by including a prior mixture weight that represents skepticism about the plausibility of an exchangeability assumption. We further stipulate a scaling factor based on the study sample sizes and intra-subject variances, when computing the predictive probabilities of bioequivalence to inform the Go/No-go decision. Once a Go decision would be made, the assessment of bioequivalence in the final analysis can be yielded by the pilot and pivotal trial data jointly. The proposed Bayesian hierarchical model is illustrated using data examples and simulations. Simulation results show that our model is superior to traditional methods, such as the two one-sided test (TOST) procedure or a modified version of the TOST.

P34

Miscellaneous

Bayesian Inference on Multilevel Semi-continuous Pharmaceutical expenditure Data

Naser Kamyari[1], Maryam Seyedtabib[2]

[1] *Abadan University of Medical Sciences, Iran*
[2] *Ahvaz Jundishapur University of Medical Sciences, Iran*

Semi-continuous data traited with an excessive proportion of zeros and right-skewed continuous positive values appear frequently in medical research. One example would be the pharmaceutical expenditure (PE) data for which a substantial proportion of subjects investigated may report zero. Two-part mixed-effects models have been developed to analyze clustered measures of semi-continuous data from multilevel studies. In this study, we propose a new flexible two-part mixed-effects model with skew distributions for nested semi-continuous cost data under the framework of a Bayesian approach. The proposed model specification consists of two mixed-effects models linked by the correlated random-effects: Part I) a model on the occurrence of positive values using a generalized logistic mixed-effects model; and Part II) a model on the magnitude of positive values using a linear mixed-effects model where the model errors follow skew distributions including beta-prime (BP). The proposed method is illustrated with a pharmaceutical expenditure data from a multilevel observational study and the analytic results are reported by comparing potential models under different random-effects structures. Simulation studies are conducted to assess the performance of the proposed models and method.

Poster Presentations

P35

Meta-analysis

Bayesian meta-analysis of predictive biomarker data from trials with partial subgroup analysis

Lorna Wheaton[1], Heather Poad[1], Sylwia Bujkiewicz[1]

[1] *University of Leicester, United Kingdom*

CONTEXT: Genetic biomarkers are increasingly used to identify subsets of a population who might benefit from therapies targeted on those biomarkers. Biomarkers are investigated in trials of heterogeneous designs, of mixed populations, such as enrichment trials investigating treatment effects in biomarker positive patients, stratified trials, or trials investigating treatment effects on all patients regardless of biomarker status. Such mixed populations introduce complexity when combining trials in meta-analysis, in particular when subgroup analyses are not conducted or reported. While individual participant data (IPD) meta-analysis can resolve this issue, IPD is rarely available for all trials. Combining aggregate data from trials of mixed populations in a single meta-analysis is important to efficiently utilise available trial data and obtain accurate and precise estimates of treatment effects for all biomarker groups.
OBJECTIVES: The objective of this study is to investigate approaches for efficient synthesis of all available aggregate data from trials with partially reported subgroup analysis in a single meta-analysis to improve estimation of treatment effects in biomarker subgroups.
METHODS: We propose using Bayesian random-effects meta-analysis, assuming systematic differences in treatment effects between biomarker groups to estimate true treatment effects for each biomarker group in each study and pooled treatment effects for each biomarker group across studies.
RESULTS: The proposed model was applied to example data in colorectal cancer where trials reported analysis of wild type (WT), mutant type (MT) and combined (WT+MT) KRAS biomarker subgroups. When only using data from trials which reported WT subgroup analysis, 8 studies were included in the meta-analysis, estimating a pooled logHR of -0.23 (SD=0.10). Using the proposed model, 16 studies (WT:8, MT:3, WT+MT:5) were included in the meta-analysis, estimating a pooled logHR of -0.25 (SD=0.054).
CONCLUSIONS: The proposed method for synthesising all available data from trials with partially reported subgroup analysis appears to reduce uncertainty around pooled estimates for biomarker subgroups compared to using random-effects meta-analysis on a reduced set of trials reporting treatment effects on a specific biomarker subgroup. A simulation study is carried out to further assess this model.

P36

Efficient clinical trial designs, Simulation studies

Bayesian methods for borrowing historical information for count data

Akalu Banbeta Tereda[1], Emmanuel Lesaffre[2], Reynaldo Martina[3], Joost van Rosmalen[4]

[1] *Uhasselt, Belgium,* [2] *KU Leuven, Belgium,* [3] *Open University, United Kingdom,* [4] *Erasmus University Medical Center, Rotterdam, Netherlands*

Including historical data into the analysis of a current clinical trial may reduce the necessary sample size and/or increase the power of analysis. However, this only applies when the historical data are similar enough to the current data. Several Bayesian approaches for borrowing historical information in a dynamic way have been proposed, such as the meta-analytic-predictive (MAP) prior and the modified power prior (MPP) both for a single historical study as well as for multiple historical studies. Here, we examine the performance of the MAP and MPP approaches for the analysis of (overdispersed) count data when multiple historical control data are incorporated into the analysis of the current data. To this end, we explore the Poisson and the negative binomial distribution.
We propose a computational approach based on path sampling for the calculation of the scaling constant of the MPP approach. We illustrate our approach using the data of a RCT involving patients with an overactive bladder whereby the response is the frequency of incontinence periods. Further, we have conducted a simulation study in case of heterogeneity of the control arms.
For similar current and historical control arms, the MPP approach offers greater statistical power than the MAP approach. When the means are different across the control arms, the MPP approach yields a slightly inflated type I error rate, whereas the MAP does not. When the dispersion parameters are different across the control arms, the results are reversed.
In conclusion, the MPP approach outperforms the MAP approach for count data.

P37

Bayesian nonparametric methods, Missing data

Bayesian nonparametric methods for prediction of missing data

Pedro Cardoso[1], TJ McKinley[1], John Dennis[1], Jack Bowden[1], MasterMind Consortium

[1] *University of Exeter, United Kingdom*

Missing data is a common problem in data analysis. Much of the literature focuses on missing outcome variables, but there are also challenges when dealing with missing predictor information, particularly when making predictions about individuals with incomplete data. This paper focuses on the development of a Bayesian model which enables an outcome prediction in the presence of missing values. Our Bayesian approach provides practitioners with probability distributions for predicted outcomes that account for any uncertainties resulting from missing predictor variables, alongside estimates for the missing information. Our method combines a regression model and a Dirichlet process mixture model (DPMM), where the former defines the treatment selection model of interest, and the latter provides a flexible way to model the relationships between the (possibly missing) predictor variables. We show that under a missing-at-random/missing-completely-at-random assumption, the DPMM can model complex relationships between variables and predict missing values conditionally on existing information. We also demonstrate that in the presence of multiple missing variables, the DPMM model can be used to identify which, if collected, would provide the most additional information about the likely outcome or whether this is necessary. We implement the new approach to improve an existing model which informs treatment selection of SGLT2-inhibitor and Dpp4-inhibitors for people with type 2 diabetes. The Bayesian framework inherently provides probability distributions for parameter estimates alongside predictions for missing values. This approach can provide practitioners with supplementary information to aid treatment selection in the presence of missing data.

P38

High dimensional data, Biomarker discovery

Benchmarking of analysis strategies for data-independent acquisition proteomics data

Eva Brombacher[1], Klemens Fröhlich[2], Oliver Schilling[2], Clemens Kreutz[1]

[1] *Institute of Medical Biometry and Statistics, Medical Center-University of Freiburg, Germany,* [2] *Institute for Surgical Pathology, Medical Center-University of Freiburg, Germany*

Making decisions can be hard. Even more so if there are an extensive number of options, as in the case of algorithms available for each step of a data-independent acquisition (DIA)-type proteomics analysis workflow.
Benchmark studies objectively compare the performance of such algorithms and, thus, can facilitate making an informed decision. In our benchmark study we evaluated more than thousand distinct data analysis workflows, i.e. different combinations of DIA software, spectral libraries, sparsity reduction, normalization, and statistical tests, which we assessed based on their ability to correctly identify differentially abundant proteins.
We found that DIA software-library combinations which include gas-phase fractionation were among the best-performing workflows, while also on average detecting the most proteins. Among all investigated statistical tests non-parametric permutation-based statistical tests consistently perform best.

Poster Presentations

P39

Personalized medicine

Benchmarking Six Variant Callers for Detecting Low Frequency Variants in Circulating Tumour DNA

Rugare Maruzani[1], Anna Fowler[1], Liam Brierley[1], Andrea Jorgensen[1]

[1] University of Liverpool, United Kingdom

Tissue biopsies are routine for informing personalised treatment options for cancer patients. However, tissue biopsies can be invasive and therefore have limited potential in disease monitoring. The utility of circulating tumour DNA (ctDNA) in blood as a non-invasive alternative has recently been investigated. The fraction of ctDNA in plasma is typically low compared to healthy cell free DNA, ultimately increasing the challenge of detecting true cancer variants. Circulating Tumour DNA mutations can occur at frequencies of <1%. Therefore, the feasibility of using ctDNA for personalising treatment depends on the ability of computational tools to detect variants at low frequencies. We benchmarked the performance of six variant callers, (Mutect2, FreeBayes, LoFreq, Octopus, Platypus and bcftools) on ctDNA datasets. We utilised a synthetic dataset to assess sensitivities at seven variant allele frequencies (VAF). As synthetic datasets may not accurately represent the complexities of real data, we also assessed the performance of callers on two real datasets, utilising a cancer mutation database in the absence of a truth set. In the synthetic dataset, LoFreq outperformed other callers, returning the best F1 Scores. Ultimately, all callers performed inadequately on synthetic ctDNA datasets. Only Mutect2 called more than 20% of true variants at 5% VAF. On average, approximately 40% of variants detected in the real dataset were classified as pathogenic. The results of this study suggest new bioinformatic tools and NGS practices are required for calling variants in ctDNA datasets. This is essential to fully realise the potential of ctDNA in personalised treatment of cancer.

P42

(Semi-)competing risks and causal inference

Causal effects of chemotherapy dose intensity on survival outcome through Marginal Structural Models

Marta Spreafico[1,2], Francesca Ieva[2], Marta Fiocco[3]

[1] Leiden University, Mathematical Institute, Netherlands, [2] Politecnico di Milano, MOX Department of Mathematics, Italy, [3] Leiden University, Leiden University Medical Center and Princes Maxima Centrum, Netherlands

Chemotherapy agents are cytotoxic, and depending on patients' toxicity levels, variations to chemotherapy intensity may be achieved by clinicians by allocating reductions or delaying the next planned dose. Therefore, negative feedback is always present between exposure to chemotherapy agents and toxicity levels, which act as time-dependent confounders. As toxicity levels are also risk factor for mortality, crude analyses give biased estimates of the causal effect of a weakened therapy-intensity on survival outcomes.
Marginal Structural Models (MSMs) [1] are a class of models that can provide unbiased estimates of the causal effect of therapy modifications. The idea is to create a pseudo-population by weighing each subject with the inverse probability of observing his/her treatment adjustment on the past toxicity levels; this technique is known as Inverse Probability of Treatment Weighting (IPTW) estimation. In such pseudo-population reductions of the chemotherapy intensity are no longer predicted by toxicities, and the effect – in both original and weighed populations – of therapy adjustments on survival can be estimated by simple crude analysis.
Building on [2] where an innovative dose-delay joint-exposure model for IPTW estimation of the causal effect of alterations to the therapy intensity is proposed, a MSM to investigate the causal effect of Received Dose Intensity (RDI) [3] on survival outcomes is developed. The focus is on the use of actual treatment data and RDI in contrast with the use of intended treatment regimen and on the incorporation of longitudinal toxicity in the model.
The new methodology is applied to the international MRC BO06/EORTC 80931 randomized clinical trial for osteosarcoma where the primary outcome was to investigate whether doses intensified regime improve survival.

REFERENCES:

[1] Hernan MA et al. Marginal structural models to estimate the joint causal effect of nonrandomized treatments. J Am Stat Assoc 2001; 96:440-448
[2] Lancia C et al. Marginal structural models with dose-delay joint-exposure for assessing variations to chemotherapy intensity. Stat Methods Med Res 2019; 28(9):2787-2801
[3] Lancia C et al. Method to measure the mismatch between target and achieved received dose intensity of chemotherapy in cancer trials: a retrospective analysis of the MRC BO06 trial in osteosarcoma. BMJ Open 2019; 9:e022980

P43

Machine learning methods for health

Causal forests for uncovering treatment effect heterogeneity and data driven subgroups in trials

Eleanor Van Vogt[1], Karla DiazOrdaz[2], Victoria Cornelius[1], Anthony Gordon[1], Suzie Cro[1]

[1] Imperial College London, United Kingdom, [2] London School of Hygiene and Tropical Medicine, United Kingdom

Traditionally, subgroup analysis in randomised controlled trials involves a statistical test for interaction between treatment and patient characteristics. However, discovery of treatment-covariate may not always lead to clinically actionable subgroups, particularly for continuous covariates. Non-parametric causal machine learning approaches are flexible alternatives for estimating heterogeneous treatment effects.
We re-analyse data from the VANISH trial, which compares early use of vasopressin with norepinephrine on renal failure free survival for patients with septic shock over 28 days. We use traditional regression methods to subgroup analysis, a data-adaptive, and a non-parametric causal machine learning method. In the traditional method, separate logistic regressions are carried out for each covariate and its interaction with the treatment on the outcome of interest. A data-adaptive method is employed using lasso regression to select the most influential interactions. Finally, we use causal forests to assess where treatment effects differ and to estimate treatment effects accurately. Causal forests comprise of honest causal trees, which use sample splitting to separately determine splits and estimate treatment effect.
The traditional logistic regression approach showed a strong interaction between serum potassium and treatment for mortality (OR 2.43, 95%CI [1.50, 4.03], p = 0.0004). No significant interactions with treatment were found when analysing renal failure. The causal forest approach identified treatment effect heterogeneity for mortality, differential forest prediction 1.17 (p = 0.06). When extracting roots splits, the most popular split was on serum potassium (mean applied threshold of 4.6 mmol/L). When dividing the patient population into subgroups based on the mean threshold, average treatment effects were 0.074 (95%CI [-0.030, 0.18]), -0.27 (95%CI [-0.39, -0.15]) for patients with serum potassium ≤4.6 and >4.6 respectively.
The causal forest agreed with other more traditional methods of subgroup analysis in identifying treatment effect heterogeneity with potassium. Whilst traditional methods may identify sources of treatment effect heterogeneity, they cannot suggest subgroups splits which are clinically actionable. The extraction of roots splits in causal forests is a novel approach to obtaining data-derived subgroups.

P44

Simulation studies, Other (if none of the topics are applicable)

Central Statistical Monitoring with False Discovery Rate Control for Multicenter Clinical Trials

Tomoyoshi Hatayama[1], Seiichi Yasui[2], Ryosuke Ura[2]

[1] Janssen Pharmaceutical K.K., Japan, [2] Tokyo University of Science, Japan

Each clinical site involved in a clinical trial needs to follow a defined operational process to ensure patient safety and the validity of the results of the studies. To ensure that clinical sites follow a defined operational process, process monitoring is important. However, 100% source data verification by on-site monitoring with visiting all sites is not considered efficient. For efficient monitoring with prioritizing the sites which should be visited, Central Statistical Monitoring (CSM) has been attracting attentions, which analyzes data centrally and detects clinical sites with suspected process abnormalities. In recent years, several methods have been proposed to compare outcomes among the sites in order to detect atypical sites with potential abnormalities on operational processes, thereby the sites to be visited are clarified. In actual clinical trials, CSM is conducted multiple times during the trials in order to detect abnormal operational processes. However, those existing methods are not considered the multiplicity of the analysis sufficiently even though the increase of false-positive rate may lead to an unnecessary frequent visits for the clinical sites. Thus, conventional approach without considering multiple comparison may not meet the CSM objectives of detecting abnormal processes efficiently.
In this study, we propose a method for multiple times CSM in multicenter clinical trials as a method of detecting abnormal clinical sites with controlling the False Discovery Rate (FDR). The proposed method is based on the control chart which originally proposed for process monitoring in manufacturing. The control chart can detect the abnormality, in which abnormality is detected when the plotted outcome in the charts exceeds the optimal boundaries which is predetermined to control the false-positive rate. In this study we analyze the outcome mean of each site in the clinical trials and the process abnormalities are detected when the mean of each site's outcome exceeds the optimal boundaries at any timepoint. The FDR can be controlled by using the optimal boundaries which are determined by Adaptive Linear Step up Procedures of Benjamin et al. (2005). The proposed method can address the multiplicity of comparison caused by frequent CSM and control the unnecessary site visits.

Poster Presentations

P45

Machine learning methods for health, Bayesian nonparametric methods

Characterising Radiological Image Quality with Heterogeneous Multi-output Gaussian Processes

Owen Thomas[1], Thien Tran[1], Peter Lauritzen[2]

[1] Akershus University Hospital, Oslo, Norway. [2] Oslo University Hospital, Norway

We present a project integrating different methods of evaluating radiological image quality within a Bayesian nonparametric framework using the Heterogeneous Multi-output Gaussian Process. Multiple measures of image quality are used, including continuous-valued physical measurements of contrast enhancement, and binary evaluations of usability from multiple clinicians: these are combined with numerous relevant covariates through kernel-based regression to provide a flexible, interpretable model that provides novel insights into the consistency of image quality rating measures and the complex clinical influence of the covariates. The challenging non-conjugate inference is performed with scalable variational methods, facilitating a greater level of statistical understanding of a complex biomedical problem. Results indicate that image quality measured exhibit an interesting correlation structure, and that contrast medium dose, time since injection and body weight are key clinical predictors of image quality.

P46

Machine learning methods for health, Personalized medicine

Cofactor-Aware Longitudinal Modelling for Neurodegenerative Disease

Nemo Fournier[1], Stanley Durrleman[1]

[1] ICM, Paris, France

OBJECTIVES: Longitudinal modelling is of pivotal interest for the study of neurodegenerative diseases. Mixed-effects approaches provide a fertile ground for this, as they disentangle population effects from individual variability. They recover both the natural history of the disease (at the cohort scale) and obtain personalized descriptors and predictions for patients. However, such methods often rely solely on assessments and biomarkers tracked over time. They thus fail to integrate time-independent cofactors (such as gender, education levels, genetic factors) in their modelling, although those might modulate the disease course. We propose a mixed-effect formulation in which clinical trajectories are influenced by such cofactors, be they categorical or continuous.
METHODS: We adapt a non-linear Bayesian mixed-effect model to explicit the dependency between time-fixed cofactors and biomarker trajectories: we learn a parametrized function that links covariates to population effects while preserving the random individual effects that capture inter-patient variability. The model thereby formulated lies in the statistical framework for which algorithms of the MCMC-SAEM family allow calibration. Given covariates and longitudinal observations, it learns how to condition the average course of the disease by any combination of those cofactors.
RESULTS: Simulation studies validated our model's ability to retrieve the effects of categorical and continuous covariates modifying the clinical evolution of a pathology (through global or feature-specific effects on onset time, progression rate or pattern). It also disentangled the influence of multiple risk factors on different sets of features while discarding covariates with no influence on the disease course.
On real cognitive assessments from an Alzheimer's Disease cohort (ADNI), we recovered clinically established effects of covariates (such as faster and earlier cognitive decline for women or APOE-ε4 mutation carriers). Future results will link finer-grained covariates (e.g., polygenic scores, SNP arrays) to a broader range of features (cognitive subdomains, MRI or PET imaging).
CONCLUSION: We describe a novel approach to model links between longitudinal trajectories and fixed-covariates, validated in simulation studies and that shows promising clinical applications.

P48

Machine learning methods for health

Combining deep learning and dynamic modelling to infer disease trajectories of patients with SMA

Maren Hackenberg[1], Michelle Pfaffenlehner[1], Astrid Pechmann[2], Janbernd Kirschner[2], Harald Binder[1]

[1] Institute of Medical Biometry and Statistics, Medical Center, University of Freiburg, Germany, [2] Department of Neuropediatrics and Muscle Disorders, Medical Center, University of Freiburg, Germany

While black-box deep learning is sometimes viewed as distinct from classical statistical modelling based on the assumption of an underlying explicit data-generating model, integrating these two paradigms allows for combining their respective advantages, and thus address complex modelling challenges for biomedical data.
Specifically, we consider neural networks for flexible dimension reduction, while we use explicit dynamic modelling to infer longitudinal patterns. We illustrate how these approaches can be combined to capture disease trajectories based on the SMArtCARE registry, a prospective cohort study of Spinal Muscular Atrophy (SMA) patients, where longitudinal data on patients' disease developments is collected during routine visits. As a particular challenge for dimension reduction, for many patients several different motor function assessments are conducted, either simultaneously or at different time points.
To address this, we employ a deep learning approach to infer a low-dimensional representation of individual developments. In a domain adaptation approach, we integrate measurements from two motor function assessments into a common latent space, using an adversarial training strategy.
In this joint latent space, we explicitly model patients' dynamics as solutions of locally defined ordinary differential equations (ODEs). Based on these solutions, we derive an unbiased estimator with minimum variance of the underlying dynamical process. Additional variables measured only at baseline are used to infer individual-specific ODE parameters.
Results show that the approach allows for inferring individual-specific disease trajectories while integrating measurements from different motor function assessments. For some patients, the two assessments cannot be completely mapped, suggesting individual differences in the degree of correspondence between them. Additionally, we are able to identify groups of patients with similar underlying dynamics, and further investigate this by jointly modelling such groups of individuals with similar trajectories to enrich the information on the common temporal dynamics.
With the proposed approach we thus more broadly exemplify the potential of combining different modelling paradigms, such as neural networks and dynamic modelling, for challenging healthcare applications.

P49

(Semi-)competing risks and causal inference, Communicating statistical methods

Combining non-adherence and mediation in a unified causal analysis

Anca Chisster[1], Richard Emsley[1], Sabine Landau[1]

[1] King's College London, United Kingdom

Clinical trials suffer from participant non-adherence. A standard intention-to-treat (ITT) analysis estimates the causal effect of treatment offer which ignores post-randomisation events such as non-adherence. Though more complex methods of analysis that account for the selection bias caused by non-adherence are required to infer causality, statistical methods for estimating the total causal effect of treatment receipt are well-established. Of these, the complier average causal effect (CACE) estimates the average effect of receiving treatment in the subgroup of participants who comply with their randomisation. Clinical trials in mental health often undertake a mechanism evaluation to understand how complex, non-pharmacological interventions, such as therapy, have led to changes in the outcome. Mediation analysis decomposes a total treatment effect into a mediated effect and a direct effect. However, current methods for mediation analysis usually decompose the ITT effect, of which its direct and mediated components ignore non-adherence. Though mediation and non-adherence are well-researched areas, it is unclear how to account for non-adherence in a mediation analysis. This paper summarises the literature on methods that combine mediation and non-adherence by considering the target estimand and its identification assumptions.
We found six papers that identify estimands for combining mediation and non-adherence; though only 3 were decomposing a randomisation preserving estimand (CACE). We show that the CACE can be decomposed into a complier average natural direct effect and a complier average causal mediated effect and estimated using linear structural equation models under a given set of assumptions. The methodology is illustrated with the AVATAR trial that evaluates the effect of AVATAR therapy versus supportive counselling on auditory verbal hallucinations. A CACE analysis shows that the average causal effect of receiving AVATAR therapy for compliers significantly decreases the outcome by 3.55 points. A mechanism evaluation found a mediation effect in those receiving AVATAR therapy through a measure of voice acceptance and action. Mediated effects for the decomposition of the CACE were similar to the decomposition of the ITT effect.
Trials should consider a unified analysis of mediation and non-adherence to accurately explore treatment mechanisms.

Poster Presentations

P50

Miscellaneous

Comparing Identifications of Average Causal Effects: Robustness and Semiparametric Efficiency

Tetiana Gorbach[1], Xavier de Luna[1], Juha Karvanen[2], Ingeborg Waernbaum[3]

[1] Umeå University, Sweden, [2] University of Jyvaskyla, Finland, [3] Uppsala University, Sweden

Semiparametric inference about average causal effects from observational data is based on assumptions yielding identification of the effects. In practice, several distinct identifying assumptions may be plausible; an analyst has to make a delicate choice between these models. In this paper, we study three identifying assumptions based on the potential outcome framework: the back-door assumption, which uses pre-treatment covariates, the front-door assumption, which uses mediators, and the two-door assumption using pre-treatment covariates and mediators simultaneously. We derive the efficient influence functions and the corresponding semiparametric efficiency bounds that hold under these assumptions, and their combinations. We compare the bounds and give conditions under which some bounds are lower than others. We also propose semiparametric estimators, quantify their efficiency and study their robustness to misspecification of the nuisance models. The theory is complemented with simulation experiments on the finite sample properties of the estimators. The results obtained are relevant for an analyst facing a choice between several plausible identifying assumptions and corresponding estimators. Here, the choice is a trade-off between efficiency and robustness to misspecification of the nuisance models.

P51

Machine learning methods for health

Comparing linear discriminant analysis and supervised learning algorithms for binary classification

Ricarda Graf[1], Sarah Friedrich[1]

[1] University of Augsburg, Germany

Linear discriminant analysis (LDA) is a widely applied method in psychology and social sciences. It comprises Fisher's DA, a descriptive method used to find the linear combination of the variables best separating two classes, and LDA used to make class predictions, for which, in contrast, the multivariate normality assumption is needed. Although it has been found that LDA may provide stable estimates of class probabilities when this assumption is violated, an insecurity remains among users. In the present simulation study, the performance of LDA is compared to those of several supervised learning algorithms, in particular Support Vector Machine (SVM) with linear kernel, Classification and Regression Tree, Random Forest (RF), k Conditional Nearest Neighbour, and Probabilistic Neural Network (PNN). As performance measures, classification indices (B, C, and Q index), measures based on the contingency table (predictive accuracy (PA), sensitivity, specificity, area under the curve (AUC)), and summary ROC curves are used.
Simulations are based on two psychological datasets, both Likert scale questionnaires, comprising five and ten variables, respectively. Two sample sizes are considered: smaller, unequal sample sizes (n0=50, n1=100) and a balanced scenario (n0=n1=500). The robustness of LDA against departures from normality is examined by applying it to datasets simulated from the multivariate normal and various nonnormal distributions.
The graphical display of the algorithms' performance in summary ROC curves indicated almost identical performance based on the larger simulated datasets. Based on the smaller datasets, RF stands out, SVM and PNN sometimes perform slightly better. It is important to notice that the summary ROC curves are extrapolated beyond the actually observed false positive rates.
In case of n0=50, n1=100, the B, C, and Q indices of LDA usually range highest or second highest for the simulations based on the two reference datasets. Values of AUC and PA are higher for the RF and SVM algorithms. In case of n0=n1=500, these differences in performance between LDA and the nonparametric algorithms SVM and RF are smaller, and sometimes reversed.
In summary, LDA is competitive in these low-dimensional psychological data using typical sample sizes, even for nonnormally distributed data. It also requires less run-time and no hyperparameter training.

P52

Simulation studies, Prediction models

Comparing Statistical and Machine Learning Approaches to Risk Prediction: A Simulation Study

Hayley Smith[1], Michael Crowther[2,3], Michael Sweeting[1,4], Paul Lambert[1,3], Tim Lucas[1]

[1] University of Leicester, United Kingdom, [2] Red Door Analytics, Stockholm, Sweden, [3] Karolinska Institute, Stockholm, Sweden, [4] AstraZeneca

A variety of machine learning (ML) approaches have been adapted to make predictions from survival data and have been compared to traditional statistical approaches in terms of predictive accuracy. However, within the current literature, there are many opposing conclusions about the comparative performance of statistical and ML approaches for risk prediction. While no difference in performance has been reported for logistic regression problems, other articles have concluded that ML approaches outperform statistical approaches for risk prediction in at least one scenario evaluated. Previous studies have been biased towards one method by using limited functional forms for their data generating mechanisms (DGMs) and therefore the results do not necessarily generalise broadly. We aimed to conduct a comprehensive simulation study to evaluate both statistical and ML approaches for a range of DGMs using a novel DGM method.
Several risk prediction methods were evaluated when estimating survival probabilities at multiple follow-up times with varying sample sizes. We compared the Cox model, Multivariable Fractional Polynomials, Flexible Parametric models, Random Survival Forests, and Cox-CC and Cox-Time neural networks. Separately for each of the methods, a DGM consisted of simulating survival times from the method using datasets derived from real data but for which some truth is known (known as a Plasmode simulation study). The aim was to highlight the performance of methods that are not the true DGM compared to the method that is the truth. Predictive and discriminative performance was evaluated using the time-dependent AUC, Brier score, and Mean Absolute Prediction Error.
We present and critique the novel simulation approach. The use of Plasmode simulations are a useful tool in evaluating prognostic models, enabling assessment of departures from assumptions for a variety of methods. We then illustrate the findings of this simulation study in terms of predictive performance, performance variability, and practical issues that can arise with these methods. Statistical methods for risk prediction were robust to different DGMs and ML method performance was highly variable across datasets. Due to increasing interest in analysing survival data with ML approaches, methods must be evaluated to ensure appropriate use and improve the accuracy of time-to-event predictions.

P53

Prediction models

Comparing subgroup-specific performance of a prognostic model using areas under weighted ROC curves

Martin Johansen[1], Mikkel Kragh[1], Jørgen Berntsen[1], Erik Parner[2]

[1] Vitrolife A/S, Denmark, [2] Aarhus University, Denmark

Evaluating and comparing the discriminatory performance of a predictor of a binary outcome is often done by means of pooled and subgroup-specific areas under the ROC curve (AUCs). But due to non-collapsibility of the AUC measure, the AUCs in subgroups may not equal the pooled AUC even when other risk factors are equally distributed between subgroups. Furthermore, the subgroup-specific AUCs might be affected by a covariate that is not distributed equally in the subgroups. One way to account for this type of heterogeneity is standardization to a common reference distribution of the covariate. However, because of the non-collapsibility, this standardization cannot be performed directly on the AUCs.
In this study, we aim to assess and compare the performance of a predictor between subgroups accounting for differences in the distribution of a covariate that is related to the performance.
We will do this using the area under a weighted ROC curve with individual-level weights for the sensitivity and specificity defined as the ratio between the covariate densities in a reference population and subgroup populations. When the reference population is chosen to be the pooled population the interpretation of the AUCs is the expected performance if the distribution of the covariate in each subgroup was the same as in the pooled population.
We illustrate the method by evaluating the performance of a predictor of hospital mortality obtained from a logistic regression model fitted on the publicly available demo of the eICU Collaborative Research Database containing information on ICU stays from hospitals across the United States in 2014-2015.
The unweighted AUCs varied between the four regions with a mean of 0.602 and a standard deviation (SD) of 0.061. However, the age distribution also varied between the four regions (mean age ranging from 58.9 to 67.0 years). Standardizing each region AUC to the pooled age distribution resulted in weighted AUCs with mean 0.623 and SD 0.036. For reference, the pooled (unweighted) AUC was 0.618.

Poster Presentations

P54

Meta-analysis

Comparison of chyle leakage rate between laparoscopic and open colectomy for right side colon cancer

Napaphat Poprom[1], Tharin Thampongsa[1], Sitanun Saengsri[1], Pichet Wattanapreechanoni[1], Chairat Supsamutchai[1], Chumpon Wilasrusmee[1]

[1] *Department of Surgery, Faculty of Medicine, Ramathibodi Hospital, Mahidol University, Bangkok, Thailand*

BACKGROUND: Laparoscopic complete mesocolon excision (LCME) for right colonic cancer improves oncological outcomes. This systematic review and meta-analysis aimed to compare rate of chylous leakage and outcomes between open or laparoscopic right colectomy with CME for right-sided colonic cancers.
METHODS: Literature searches of electronic databases and manual searches up to February 1, 2022, were performed. Primary outcome was the rate of chylous leakage. Secondary outcomes were anastomosis leakage, blood loss, harvested lymph nodes number, overall morbidity, and operative time. A meta-analysis was performed to calculate risk ratios.
RESULTS: Eleven studies were identified with 1,542 patients enrolled. The rate of postoperative chyle leakage in laparoscopic had lower than open surgery with pooled RRs of (MD = 0.63, 95% CI; 0.33, 1.20) although these was not significant. LCME was significantly better than OCME in terms of blood loss (MD = -0.49, 95% CI; -0.83, -0.15), harvested lymph nodes number (MD = 0.06, 95% CI; - 0.24, 0.37) and overall morbidity (RR = 0.59, 95% CI; 0.43, 0.80). There was no significant difference regarding anastomosis leakage, and operative time.
CONCLUSIONS: The results demonstrate that the rate of chylous leakage in LCME is not different compared to OCME. LCME in right colon cancer surgery is superior to OCME in terms of blood loss, harvested lymph nodes number and overall morbidity.

P55

Simulation studies

Comparison of different propensity score methods in a simulation study

Arianna Galotta[1], Alice Bonomi[1], Giancarlo Marenzi[1]

[1] *Centro Cardiologico Monzino, Milan, Italy*

INTRODUCTION: In observational studies, the causal association between event and treatment can be altered by the presence of bias, confounding, and possible imbalances between treatment groups. There is growing interest in the literature in the use of propensity score (PS) to overcome this bias.
OBJECTIVE: Apply and compare the different applications of PS methods (PS matching, stratification, inverse probability of treatment weighting, overlap weighting and regression using PS as an adjustment variable) in a simulation study.
METHODS: The simulated dataset is generated from the information obtained from the observational data: the distributions of the variables were approximated by the empirical distributions of the confounders and the treatment and/or outcome-related variables. Different conditions in according to different applications of PS are created, varying the sample size and varying the intensity of effect of treatment and confounders on outcome. PS application methods are compared in terms of variance, bias, and mean square error (MSE) of the estimator.
RESULTS: Regression with PS as an adjustment variable and regression with single variables perform better. In addition, the worst performance in terms of MSE in all created scenarios is associated with PS matching, this due to higher variance of the estimates. Sample size obviously has an impact on variance but not on bias: as the size increases, the variance decreases. The introduction of variables related only to treatment increases the variance of the estimator.
CONCLUSIONS: In conclusion, PS methods can be used to control for confounding in observational studies. The use of PS should be evaluated based on the available data. In this simulation study, PS matching is the least optimal method to use; whereas regression with PS or classical regression with single covariates perform better.

P56

Meta-analysis

Comparison of methods for partial reporting of multiple thresholds in test accuracy meta-analysis

Matthew Bond[1], Joie Ensor[1]

[1] *Keele University, Newcastle, United Kingdom*

BACKGROUND: Commonly studies examining the accuracy of a test based on a continuous diagnostic variable, report test performance at multiple thresholds. When systematic reviews synthesise this evidence from multiple studies, we commonly observe heterogeneity in the thresholds chosen across the included studies, leading to 'missing' threshold information at meta-analysis. This missingness is likely due to selective reporting of threshold information, leading to bias and a lack of precision in summary meta-analysis estimates on test performance. In recent years several novel methodologies have been proposed to tackle this issue.
OBJECTIVE: We aimed to systematically review the literature to identify existing methods addressing this issue and to subsequently assess the performance of these methods across a range of scenarios to identify the most promising methods.
METHODS: Bibliographic databases were searched for index terms relating to diagnosis, meta-analysis and thresholds. Screening of titles, abstracts and subsequently full texts was conducted by two reviewers independently using predefined criteria. Following the systematic review, we conducted an extensive simulation study designed following the ADEMP structure for planning and reporting simulation studies.
Complete individual participant data from primary studies was simulated before generating missing threshold information assuming different missingness mechanisms. We investigated a range of scenarios by varying parameters of the simulation including, the number of studies, number of patients, disease prevalence and diseased/non-diseased distributions. Performance measures included bias, mean squared error, coverage and convergence.
RESULTS: We identified a total of 9293 unique records and found ten studies developing novel methods to handle partial reporting of multiple thresholds in test accuracy meta-analysis. Methodological approaches included, estimating the underlying distribution of the diagnostic variable, and imputing missing threshold information based on reported threshold data.
CONCLUSIONS: Our simulations indicate that all methods compared in this simulation outperform the standard meta-analysis approach which excludes primary studies not reporting the threshold of interest. No single method was found to outperform the others across all considered scenarios.

P57

Communicating statistical methods

Comparison of statistical methods for the analysis of SF-36 in RCTs: an empirical analysis

Yirui Qian[1], Stephen Walters[1], Richard Jacques[1]

[1] *University of Sheffield, United Kingdom*

BACKGROUND: Short-Form 36 (SF-36), a patient-reported outcome measure, is frequently used as a key outcome in randomised controlled trials (RCTs). The SF-36 generates eight health domain scores from 36 items on ordinal scales. The domain scores tend to be discrete, bounded, and skewed. An inappropriate analysis can result in unreliable estimates, with wider confidence intervals (CIs) or larger errors. This study aims to compare the different statistical methods for the analysis of SF-36 domain scores using empirical analysis.
Method: Ten statistical methods included for comparison were linear regression, tobit model, median regression, censored absolute least deviation (CLAD) model, ordinal regression (ordered logit & ordered probit model), binomial regression (beta-binomial & binomial-logit-normal regression), and fractional regression (beta regression & fractional logistic regression). Data analysis was conducted by fitting these methods to ten RCT datasets, which used the SF-36, in multiple clinical areas.
RESULTS: Marginal effects of the tobit model produced similar estimates of treatment effect as linear regression, while estimates by median regression and CLAD model deviated from the estimates by linear regression. When the number of possible scores is less than five, the CLAD model did not always converge. Two pairs of methods - ordered logit & ordered probit model and beta-binomial & binomial-logit-normal regression had similar model performance. Ordered logit generated higher absolute estimates than other statistical methods with logit link function. When the number of possible scores of the SF-36 domain increased, the model fit of tobit regression, ordinal regression and binomial regression became poorer, but the model fit of linear regression improved.
DISCUSSION: Linear regression is a classic and simple method to use when model assumptions are not significantly violated. Ordinal, binomial, and fractional regression may fit better to domains with small number of possible scores, but they require recoding of SF-36 scores and transformation of estimates for interpretation, which increases the complexity in practice. To what extent the increase in the model fit outweighs the cost of simplicity of interpretation needs to be investigated. Future research will focus on simulation analysis to compare the estimation accuracy of these methods.

Poster Presentations

P58

Communicating statistical methods, Statistical education

Comparison of survival models for non-proportional hazards: a cohort study using the UKTR

Jyoti Sehjal[1]

[1] *London School of Hygiene and Tropical Medicine, United Kingdom*

Non-proportional hazards are common in the context of all-cause mortality after surgical interventions where specific risk factors have different effects on the competing causes of death. The hazard function for death after liver transplantation (LT) has a "bathtub" shape in that there is initially a high hazard of death from the surgery, primary graft non-function and infection, followed by a period of low hazard, before the hazard increases again due to other issues, such as cancer recurrence or other comorbidities. Though some parametric models can model these non-monotonic hazards, they are not widely used in the medical literature. The Cox proportional hazards model remains the most popular method for analysing time-to-event data, mainly because it does not require specification of the baseline hazard function. However, the advantages of parametric survival models are they are more efficient than semi-parametric alternatives if the functional form of the baseline hazard function is known, and estimation of the survivor function and covariates is faster if the hazard function can be integrated analytically. Moreover, they give more precise predictions of long-term survival provided extrapolations are reliable.
This presentation aims to compare the merits and pitfalls of parametric, flexible parametric and semi-parametric methods available to analyse time-to-event data that exhibit non-proportional hazard functions and guide the choice of which model to use in different scenarios. Among these models are the extended Cox proportional hazards model using restricted cubic splines to flexibly model time-varying associations between covariates and survival, the Royston-Parmar flexible parametric survival model, and the parametric survival model based on the generalised gamma distribution. Models were applied to study survival after LT using a cohort of 6,784 adult recipients from the United Kingdom (UK) who received their first elective orthotopic liver transplant between 1997 and 2016. Comparisons of model fit are made using Akaike and Bayesian information criteria and comparisons of survival used estimated survival curves and relative hazards.

P59

Efficient clinical trial designs, Simulation studies

Comparison of the KEYBOARD and BOIN designs for phase I trial sensitivity to cohort deviations

Joshua Roche[1], Helene Thygesen[1]

[1] *University of Glasgow, United Kingdom*

The traditional 3+3 design, used in many dose finding studies, requires patients to be recruited in "cohorts" of a fixed size as the dose escalation rule requires the toxicity status for a specific number of patients to be known. One of the advantages of model-based designs such as the Continual Reassessment Method (CRM), the KEYBOARD design and Bayesian Optimal Intervals (BOIN) is that the models can prescribe a Next Recommended Dose for any history of toxicity reports. However, we typically still need to pre-specify the cohort size even for these designs, otherwise it is difficult to evaluate their operating characteristics.
Unfortunately, deviations in cohort size pervade phase I trials due to problems such as slow recruitment, loss to follow-up and patients being unevaluable; exacerbated further recently by the COVID pandemic. In the case of the 3+3 design, this can cause longer trial durations and hence delay the subsequent phase II/III trials. However, since the predicted operating characteristics of the model-based designs assume an optimal scheduling of patient arrivals (as the design is most efficient if patients are recruited at the time at which enough information has accumulated to revise the next recommended dose), the association of these deviations with the operating characteristics of the model-based designs can be unclear.
In this study, we simulated a number of trials under varying degree of cohort deviation complexity and evaluated the KEYBOARD and BOIN designs in terms of key metrics including trial duration, number of patients required, over- and underdosing risk and statistical power. Preliminary results indicate both the BOIN and KEYBOARD designs are largely comparable and that if the cohort size change is independent of the observed data, there is very little impact on the operative characteristics of either design.

P60

Machine learning methods for health, Prediction models

Comparisons between traditional and machine learning approaches to clinical prediction models

Haoyue Wang[1]

[1] *University of Manchester, United Kingdom*

Clinical prediction models (CPMs) are frequently developed to predict risks of outcomes for patients. For example, the Trauma Audit and Research Network (TARN) predicts mortality for major trauma patients in England and Wales and their comparisons of predicted and observed outcomes have been used to inform NHS-wide reorganisation of care. Currently, TARN is investigating the prediction of quality of survival for patients after 6 months' recovery. The patient-reported outcome measures of interest include binary, ordinal and continuous variables. The predictors include patient characteristics and injury details of interest. To predict the survival quality outcomes, both traditional statistical methods (e.g., logistic regression, ordinal regression) and machine learning approaches (e.g., neural network, random forest) can be used. Each of them, however, has their own strengths and limitations. Therefore, it is difficult to identify a single method which uniformly outperforms the other ones. Indeed, comparisons between methods may not be straightforward or even meaningful. It would be helpful to understand the similarities and differences between these approaches such that appropriate CPMs can be developed for different research questions and outcome variables.
In my project, I investigate statistical and machine learning approaches to predicting the quality of survival using data from TARN. In particular, I evaluate how current methodological and reporting guidance applies to different modelling approaches. I explore the sample size required in different CPM development approaches and consider the impact of sample size on the model validity.

P61

Miscellaneous

Correlated Random Effects in Generalized Linear Mixed Models Associated with Designed Experiments

Zakir Hossain[1], Heiko Grossmann[2], Steven Gilmour[3]

[1] *University of Dhaka, Bangladesh,* [2] *University of Magdeburg, Germany,* [3] *King's College London, United Kingdom*

In generalized linear mixed models (GLMMs), random effects are usually uncorrelated and normally distributed. We extend the idea of randomization for deriving linear models to GLMMs using different permutation groups. We derived three GLMMs and corresponding variance-covariance matrices for the random effects using the symmetric group for completely randomized design (CRD), the wreath product for randomized complete block design (RCBD) with random block effects; and the direct product for RCBD with fixed block effects. In our case, random effects are correlated due to randomization. Simulation studies have been conducted for estimating the model parameters considering the Poisson regression mixed model.

Poster Presentations

P62

Biomarker discovery

Correlates of Protection Analysis of Vaccines in a Pre-Exposed Population

Joris Menten[1], Els De Paepe[1], Efi Gymnopoulou[1], Roels Sanne[1], Arangassery Rosemary Bastian[1]

[1] *Janssen Pharmaceutical Companies of Johnson & Johnson, Belgium*

CONTEXT: Correlates of protection (CoP) are biomarkers that are statistically correlated with and predictive of vaccine efficacy. They are important in vaccine development to predict durability of protection and select the optimal dose and regimen.
Statistical theory for CoP analyses, often within a causal inference framework, is typically developed for a population that had no previous exposure to the pathogen of interest. However, in the development of a vaccine against the Respiratory Syncytial Virus (RSV) in older adults, the population of interest is by the time of vaccination already pre-exposed to RSV. This leads to additional complexities in assessing whether a biomarker is a CoP due to the variation in pre-existing biomarker levels.
OBJECTIVES: Illustrate the complexities in CoP analyses when applied to a pre-exposed population and present some refinements to existing methods.
METHODS: Using simulation studies, we demonstrate the characteristics of standard and refined CoP analyses. More specifically, antibody (Ab) data, predictive of protection against RSV, after vaccination, are simulated under different levels of pre-existing immunity. This pre-existing immunity may depend on many factors: the number of and the time since previous RSV infections, immuno-senescence, and comorbidities.
RESULTS: The (rise in) levels of antibodies immediately after vaccination may be used as CoP, but both pre- and post-vaccination Ab levels are influenced by many interacting factors which confound the relationship between vaccine-induced immune response and protection from disease. Even though a vaccine may protect by increasing the Ab levels, an analysis of the fold rise in Ab levels following vaccination may be strongly influenced by variation in pre-existing immunity. Results of analyses of the fold-rise in Ab levels as a CoP can deviate from the true underlying effect of increasing Ab levels through vaccination. In addition, prediction of the durability of vaccine efficacy through peak Ab levels after vaccination is complicated by the decay in Ab levels from the peak response in vaccinated participants and from previous infection in unvaccinated.
CONCLUSION: CoP analyses of vaccines in a pre-exposed population must take into account levels of pre-existing immunity and the correlation between pre- and post-vaccination Ab levels.

P63

Communicating statistical methods, Personalized medicine

Cumulative Sum Monitoring of Arbitrarily Censored Time-to-Event Processes

Mireya Diaz[1]

[1] *Homer Stryker M.D. School of Medicine, Western Michigan University, Kalamazoo MI, United States*

BACKGROUND: Statistical process control has made headway in health care as part of continuous process improvement leading to considerable methodological developments in monitoring based on time-to-event cumulative sum (CUSUM) charts in the past fourteen years. Extensions to the initial chart include risk-adjustment, dependent censoring, long-term survivorship, estimation errors, and shared frailties. The extension presented here considers events, which are interval-censored and handle them together with exact and right-censored events. Examples of these intervals include processes with unknown exact event times, e.g. those that occur between visits, as well as patient reported outcomes subject to recall bias.
METHODS: Formulae for the CUSUM and learning curve (LC) CUSUM weights were developed to include interval censoring providing a closed-form solution for their estimation. These weights signal processes that reach the out-of-control and the in-control state respectively. These formulae where examined with empirical data to gain insights as to the role of several factors on the chart weights.
RESULTS: The extension to handle interval-censoring events from existing CUSUM and LC-CUSUM weights is straightforward. Interval-censored observations contribute to the CUSUM scores with weights that include a ratio of parameters from the alternative and null hypotheses within exponential differences of the interval boundaries combined with parameters of the survival distribution, for example Weibull. Empirical analysis shows that the initial interval time has a much greater effect on weights than the interval width. Interestingly, although resulting in a small error, mid interval imputation is not necessarily the one generating the smallest difference, potentially because the median of the process, based on the Turnbull estimator, is closer to the left end of the intervals in the example examined.
DISCUSSION: A closed-form formulae for the weights of CUSUM and LC-CUSUM charts for time-to-event outcomes were developed. Empirical analysis drew initial information about features influential to the weights. This will focus a subsequent study to evaluate the role of factors such as size of the in-control and out-of-control series, parameters of the event time distribution, interval starting time, percent censoring, and alarm thresholds.

P64

Efficient clinical trial designs

Data-Dependent Contrast Test for Dose-Response Studies

Masahiro Kojima[1,2]

[1] *Kyowa Kirin Co., Ltd, Japan,* [2] *The Institute of Statistical Mathematics, Tokyo, Japan*

A confirmation of dose-responses is complicated by adjustments for multiplicity. We We propose a simple and powerful data-dependent contrast test with ordinal constraint contrast coefficients determined by observed responses. The contrast coefficients are easily calculated by combining a pool-adjacent-violators algorithm and assumptions of the contrast coefficients. The data-dependent contrast test can be performed using SAS, and we provide the sample SAS program code of analysis and power calculation for the data-dependent contrast test. After the data-dependent contrast test shows the statistically significant dose-response curve, the best dose-response model can be selected from multiple dose-response models. Based on the best model, a recommended dose can be identified. We demonstrate the data-dependent contrast test for sample data. In addition, we calculate the ordinal constraint contrast coefficients and test statistic for the actual study, and we show a recommended dose. Finally, we perform the simulation study with eleven scenarios to evaluate the performance of the data-dependent contrast test by comparing multiple comparison procedures with modeling techniques. We confirmed the statistically significant dose-response curve for both the sample data and the actual study. In the simulation study, the data-dependent contrast test had higher power for all scenarios than the conventional method. In addition, the type I error rate of the data-dependent contrast test was maintained at a significance level when there was no difference between the treatment groups. We conclude that the data-dependent contrast test can be applied unproblematically to the dose-response study.

P65

Personalized medicine, Other (if none of the topics are applicable)

Developing a Patient Engagement Framework for Quality Improvement in Healthcare Services

Roma Puronaite[1], Justas Trinkunas[1], Andrejus Cernovas[1]

[1] *Vilnius University Hospital Santaros Klinikos, Lithuania*

Ensuring coherence between the patient's quality of life and the patient's experience of the healthcare system, identifying and removing existing barriers to the organisation of care, and achieving optimal health outcomes for patients are key to improving the quality of care. A comprehensive, patient-centred healthcare process analysis tool can improve the quality management of healthcare services.
Our project aims to develop a new, original and integrated patient engagement framework, compatible with existing eHealth and national healthcare systems. To achieve this objective, qualitative and quantitative research is carried out.
The development process consists of the following steps: (1) analysis of international experience and research on measures of patient-reported outcomes and patient-reported experiences, (2) focus groups to analyse the patient journey in the healthcare system and to identify potential barriers for patients, (3) a representative survey of the population, (4) the development of the patient engagement framework as an integrated tool for collecting, processing and analysing the patient-reported experience, and (5) a field study to test the tool in patient groups.
The main outcome of the project will be the development and testing of a new scientific product, a new tool for patient involvement in the healthcare process, and the analysis of patient reported data.

Poster Presentations

P66

Efficient clinical trial designs, Personalized medicine

Development of a generalisable methodology for n-of-1 trials – results from a review of trials

Robin Chatters[1], Olivia Hawksworth[1], Steven Julious[1]

[1] *University of Sheffield, United Kingdom*

BACKGROUND: Patients can respond in different ways to the same treatment depending on factors including age, genetics and lifestyle. Therefore, 'one size fits all' may be an inappropriate philosophy when it comes to making treatment decisions. n-of-1 trials are a type of multi-period crossover study in which the effects of a treatment are investigated by following an individual patient over time, where the treatments given are randomised from period to period. The review aimed to identify characteristics of n-of-1 trials in rare and non-rare diseases, including their design, to facilitate recommendations for future n-of-1 trials.
METHODS: Clinical trial registries and electronic databases of published journal articles were searched using the term "n-of-1", filtering for articles published between 2011 and May 2021. The resulting articles were screened; any that did not report on an n-of-1 trial in humans, and did not present individual treatment level data, were excluded. Data were extracted from the included studies relating to the study itself, the treatment and the outcomes used.
RESULTS: 243 studies were identified from the searches, with 53 being included in the review. Most studies assessed pharmaceutical treatments (n=35, 66.0%). Of the studies, 64.2% were placebo controlled (n=34) and 24.5% employed an active control (n=13). Only eight studies (15.1%) were undertaken in rare diseases. The median number of treatments compared was two, across a median of six periods. Eleven studies (20.8%) were conducted in a single patient only. The mean number of patients enrolled in (in multiple n-of-1) trials was 14 (maximum = 60) and of the studies conducting multiple n-of-1 trials (n=42), 78.6% (n=33) undertook, or planned to undertake, a meta-analysis to estimate group level effects. Patient retention in the studies was high, at 82.6%. Characteristics of n-of-1 trials in rare and non-rare disease areas were consistent.
DISCUSSION: A typical n-of-1 trial consists of six periods and evaluated two treatments. The results from this study may be used by researchers to aid the design of future n-of-1 trials. The DIAMOND (development of generalisable methodology for n-of-1 trials delivery for very low volume treatments) project which this work informs will give recommendations for n-of-1 trials.

P67

Prediction models

Development of a prognostic model for quality of life decline in head and neck cancer survivors

Erlend Fossen[1], Mauricio Moreira Soares[1], Andrew R. Ness[2], Steve Thomas[3], Lisa Licitra[4], Arnoldo Frigessi[1], Marissa LeBlanc[1]

[1] *University of Oslo, Norway,* [2] *University Hospitals Bristol, United Kingdom,* [3] *University of Bristol, United Kingdom,* [4] *Università di Studi di Milano, Italy*

Head and neck cancer (HNC) survivors are defined as those patients that have survived their initial treatment with no detected recurrence. However, side effects from treatment can persist for years and impact quality of life (QoL), such as living with the fear of recurrence and long-term disability. Thus, the aim of this study is to develop a prognostic model and identify variables that predict self-reported QoL 3 years post-diagnosis using data from 3011 HNC survivors from the Head and Neck 5000 study. The outcome of interest is a clinically-relevant decline in QoL over 24 months (from end of treatment to last follow-up). It is not possible to model this without conditioning on survival to last follow-up, therefore we model the probability of QoL decline given covariates and survival to last follow-up. To incorporate that our measure of QoL decline is binary, but dependent on individual survival, the prediction model was split into two parts. The aim of the model was to predict the probability of being alive and having a decline in QoL, given a set of covariates. This probability was calculated as the product of: a) the probability of having a decline in QoL given covariates and being alive; b) the probability of being alive given covariates. The data were split, stratified by outcome, into training and test data using a 80/20 split. A random forest model with 292 predictors was used. The predictors included demographic and clinical information at diagnosis, as well as psychometrics and quality of life scales obtained at several time points (diagnosis, 4 months and 12 months after diagnosis). Using Shapley values and other metrics we will present the predictors that are most important for predicting QoL decline and the interactions between them. In addition, general prediction performance of the model will be assessed with AUROC. We want to emphasise the importance of taking survival into account when predicting outcomes that are linked to survival.

P68

Machine learning methods for health

Diagnosis (ICD-10) prediction from discharge summary by Deep Learning

Wanchana Ponthongmak[1], Nawanan Theera-Ampornpunt[1], Ratchainant Thammasudjarit[1], Oraluck Pattanaprateep[1], Ammarin Thakkinstian[1]

[1] *Mahidol University, Bangkok, Thailand*

Despite various benefits (i.e., inpatient care, healthcare management and research, and reimbursement) of assigning the ICD-10 code for diagnosis, the ICD-10 code assignment is challenging, which requires an understanding of the medical domain and the ICD coding structure. In practice, most clinical data are written in free text by clinical practitioners, and medical coders perform the ICD-10 code assignment. Such coding practice leads to coding errors. Automate ICD-10 coding tool developed from deep learning for natural language processing (NLP) might reduce the burden on medical coders, which could reduce coding errors. This study initiated the development of the NLP model to assign ICD-10 codes from clinical textual data by predicting the top fifty ICD-10 codes from clinical notes in Ramathibodi Hospital's discharge summary. 17,758 eligible admissions were retrieved from the Ramathibodi hospital information system from 1st January 2015 to 31st December 2020. Three models were developed to compare model performance, including 1) Naïve Bayes model with term frequency inverse document frequency (TF-IDF), which was a baseline model, 2) deep learning model convolutional neural network (CNN) with neural word embedding, and 3) deep learning model using CNN with PubMed bidirectional encoder representations from transformers (PuMedBERT). The dataset was split into training, validation, and test datasets. The training set was used to develop models, the validation set was used for fine-tuning hyperparameters, and the test set was used to evaluate model performance and generalization. The results showed that the deep learning model with PuMedBERT provided the best performance with micro and macro average ROC of 0.9515 (95%CI 0.9440 - 0.9590) and 0.9211 (95%CI 0.9117 - 0.9306), respectively. On the other hand, the performance of the deep learning model with neural was embedding were 0.9446 (95%CI 0.9366 - 0.9526) and 0.9187 (95%CI 0.9092 - 0.9283), while the baseline model was 0.8855 (95%CI 0.8743 - 0.8966), and 0.8459 (95%CI 0.8333 - 0.8585), respectively. The results indicated that the deep learning for NLP outperformed the traditional machine learning for NLP with less afford on the feature extraction.

P69

Prediction models, Personalized medicine

Dichotomisation of continuous outcomes in prediction model research: a review of current practice

Lucinda Archer[1], Miriam Hattle[1], Kym IE Snell[1], Joie Ensor[1], John Allotey[2], Shakila Thangaratinam[2], Richard Riley[1]

[1] *Keele University, Newcastle, United Kingdom,* [2] *University of Birmingham, United Kingdom*

OBJECTIVES: (i) to review and appraise how continuous outcomes are being modelled in the development of prediction models, (ii) to summarise the justification given by researchers for any categorisation of continuous outcome variables, and (iii) to provide examples of the impact of categorisation over modelling on the continuous scale.
DESIGN: Systematic review of the literature, to identify prediction models for the continuous outcome of birthweight (as an indicator of fetal growth restriction [FGR]).
STUDY SELECTION: Studies developing a prediction model related to birthweight or FGR (by any definition) in a population of pregnant women. Studies were required to have fully reported their model equation, including coefficient and intercept terms, and to contain at least three predictor variables. Reports on external validation only, without model updating, were excluded.
DATA EXTRACTION: Two authors independently extracted data, using a pre-designed and piloted extraction form. Data included information on the handling of the birthweight outcome, the basis used to define any cut-points of any outcome categorisation, and any justification authors gave for their outcome modelling choice.
RESULTS: Database searches yielded 48 relevant papers, published between 1985 and 2019, containing 99 prediction models. Only 12 (12%) of these models aimed to predict birthweight on the continuous scale, five of which came from the same publication. Of those papers using a logistic regression framework with a dichotomised outcome, eight predicted the risk of FGR including some dichotomy of birthweight with associated complications, ten predicted the risk of being small-for-gestational-age (SGA) based on a simple dichotomisation of birthweight at established population centiles. Justification for the choice of cut-points was often lacking and inconsistent across studies. In a case study, we demonstrate the advantages of modelling on the continuous scale, and how dichotomisation can be applied at the end of the modelling process.
CONCLUSIONS: Outcome dichotomisation is common in the prediction of birthweight, often without sound justification, and using varied cut-points across studies. We recommend that continuous outcomes are rather modelled on their continuous scale, to allow predictions of continuous values that retain a more complete picture for informing decisions.

Poster Presentations

P70 Miscellaneous

Diet, Nutrition, Obesity, and Their Implications for COVID-19 Mortality: A Marginalized 2-Part Model

Naser Kamyari[1], Maryam Seyedtabib[2]

[1] *Abadan University of Medical Sciences, Iran*
[2] *Ahvaz Jundishapur University of Medical Sciences, Iran*

BACKGROUND: Nutrition is not a treatment for COVID-19, but it is a modifiable contributor to the development of chronic disease, which is highly associated with COVID-19 severe illness and deaths. A well-balanced diet and healthy patterns of eating strengthen the immune system, improve immunometabolism, and reduce the risk of chronic disease and infectious diseases.
METHODS: We globally evaluated the distribution of diet and nutrition at the national level while considering the variations between different World Health Organization regions. The effects of food supply categories and obesity on (as well as associations with) the number of deaths and the number of recoveries were reported globally by estimating coefficients and conducting color maps.
RESULTS: The findings show that a 1% increase in supplementation of pulses reduced the odds of having a zero death by 4-fold (OR 4.12, 95% CI 11.97-1.42). In addition, a 1% increase in supplementation of animal products and meat increased the odds of having a zero death by 1.076-fold (OR 1.076, 95% CI 1.01-1.15) and 1.13-fold (OR 1.13, 95% CI 1.0-1.28), respectively. Tree nuts reduced the odds of having a zero death, and vegetables increased the number of deaths. Globally, the results also showed that populations (countries) who consume more eggs, cereals excluding beer, spices, and stimulants had the greatest impact on the recovery of patients with COVID-19. In addition, populations that consume more meat, vegetal products, sugar and sweeteners, sugar crops, animal fats, and animal products were associated with more death and less recoveries in patients. The effect of consuming sugar products on mortality was considerable, and obesity has affected increased death rates and reduced recovery rates.
CONCLUSIONS: Although there are differences in dietary patterns, overall, unbalanced diets are a health threat across the world and not only affect death rates but also the quality of life. To achieve the best results in preventing nutrition-related pandemic diseases, strategies and policies should fully recognize the essential role of both diet and obesity in determining good nutrition and optimal health. Policies and programs must address the need for change at the individual level and make modifications in society and the environment to make healthier choices accessible and preferable.

P71 Simulation studies, High dimensional data

Disagreement Based Variable Selection Method for High-dimensional Censored Data

Suborna Sultana[1], Md. Hasinur Rahaman Khan[1]

[1] *Dhaka University, Bangladesh*

The sparsity associated with high dimensional censored data becomes more complex, because of the incompleteness due to censoring. In such case, survival data can not be handled by standard linear regression techniques. In this study, we use regularized linear regression model with weighted least-squares to handle the survival prediction in the presence of censored instances in high-dimensional data. Very recently in 2021, Yuen and Fryzlewicz propose a Combined Selection and Uncertainty Visualizer (CSUV), which visualises selection uncertainties for covariates in high-dimensional linear regression by exploiting the disagreement among different base selectors. Their method highlights covariates that get selected the most frequently by the different base variable selection methods on subsampled data. In this study, we extend Yuen and Fryzlewic's proposed method in the context of high-dimensional censored data through the accelerated failure time (AFT) model framework. Performance of the extended method have been examined with several simulation studies conducted under variety of settings including different collinearity level among the covariates individually or group-wise and censoring levels. Moreover, a real micro-array data example on Diffuse Large B-cell Lymphoma (DLBCL) patients has been performed to demonstrate the performance of the proposed methods for identifying correct genes that are related with the survival time of the DLBCL patients. Simulation along with the real data analysis results suggest that the extended CSUV method for censored data perform considerably well compared to some model selection procedure like extended BIC (eBIC) and delete-n/2 cross validation (CV) in most of the cases. It also outperforms the traditional regularized methods like Lasso, Elastic Net, relaxed Lasso, MCP and SCAD.

P72 Machine learning methods for health

Discrimination and calibration of atrial fibrillation predictions obtained by CNN and XGB from ECG

Giulia Barbati[1], Arjuna Scagnetto[2], Giovanni Baj[1], Luca Bortolussi[1]

[1] *University of Trieste, Italy,* [2] *Cardiovascular Department, Azienda Sanitaria Universitaria Giuliano Isontina (ASUGI), Trieste, Italy*

BACKGROUND: Atrial Fibrillation (AF) is a heart arrhythmia which, despite often asymptomatic, represents an important risk factor for stroke. The only way to diagnose AF is by an electrocardiogram (ECG) exam. Being able to predict the future insurgence of AF at the ECG exam would be of great impact on actively targeting patients at high risk.
OBJECTIVE: The focus of the study is to investigate variations in predictive performance according to different combinations of balanced/unbalanced training and test sets and sample sizes. The outcome is the prediction of new onset AF using ECG signals with a maximal distance to AF development of 5 years.
METHODS: Algorithms under study are Convolutional Neural Networks (CNN), eXtreme Gradient Boosting (XGB) and a logistic regression model that integrates CNN and XGB results (LRint). As performance measures we use the area under the ROC curve (AUC), accuracy, sensitivity, specificity, calibration intercept and slope. ECG were recorded during routine exams, at the Cardiovascular Department of Azienda Sanitaria Universitaria Giuliano Isontina (ASUGI) in Trieste, Italy. Outcome data come from the integration of ECG recordings within the regional data warehouse that collects all administrative healthcare database, such as hospitalizations, demographics, drugs prescriptions, laboratory exams and the cardiological clinical evaluations. In the cohort construction, a maximal distance of 5 years between the ECG exam and AF or censoring date (31 December 2020) was allowed.
RESULTS & CONCLUSIONS: Starting from the available population, patients with AF already diagnosed before 2005 or implanted with rhythm-controller device any time during the observation period have been excluded. The database counts 92.557 patients and 207.748 ECG, in a time-window of 15 years, from 2005 to 2020. The AF incidence is 11% (10.078 subjects/25.885 ECG signals). The dataset selected for the study has dimension ranging from 2.000 to 20.000 subjects as total sample size (TTS), with increasing step of 2000 subjects. Training and test set are built on 80/20 proportion of the TTS. The training sets are kept always balanced, while the test sets are balanced (50% AF vs 50% not-AF) and unbalanced as the target population proportion (11% AF). We are currently elaborating the results about the performance measures.

P74 Simulation studies, Statistical education

Effect of Parametrizations and Transformations on Hierarchical Models using Hamiltonian Monte Carlo

Aglina Lika[1], Dimitris Rizopoulos[1], Meng Yuan[2], Michelle Kruijshaar[1], Eleni-Rosalina Andrinopoulou[1]

[1] *Erasmus Medical Center, Rotterdam, Netherlands,* [2] *University Hospitals Leuven, Netherlands*

In medical fields, large amounts of complex data, which often form clusters or groups with familiar properties, are generated daily to monitor the progression of chronic diseases. For example, in Pompe disease, the interest lies in investigating the evolution over time of clinical but also patient reported outcomes after treatment initiation with Enzyme replacement therapy in the Erasmus MC in the Netherlands. A popular framework for analyzing such data is hierarchical modelling, also known as multi-level or mixed-effects modelling. Under the Bayesian framework, these models can be computationally expensive and therefore fail to converge. Such situations typically appear in applications where non-linear, high-dimensional parameter space and complicated random effects structures are assumed. Several solutions to improve computational efficiency have been examined. Previous work focused on either investigating different parameterizations in hierarchical models or investigating different ways of adjusting the magnitude of the variables using Markov chain Monte Carlo (MCMC) algorithms. However, little work has been done on whether these frameworks could be incorporated together to improve the performance of the models. We aim to improve the efficiency of the hierarchical model by putting into a common framework the use of data transformations and parameterizations and to investigate in which settings, which combinations of these methods are more optimal when assuming the Hamiltonian Monte Carlo algorithm. Also, we go beyond the standard centering transformation and investigate whether standardization could improve the model's efficiency. Using an extensive simulation study we observed that, most of the times, the hierarchical centering parametrization (HCP) is better than the hierarchical non-centering parameterization (HNCP). The models with centered data perform the same as the models with standardized data. The same goes for the HNCP models. Similar results are obtained from the application. In the application the HNC models with centered data were the best.

Poster Presentations

P76

Communicating statistical methods, Simulation studies

Estimating cumulative protracted time-lagged effects of training load on injury: a simulation study

Lena Bache-Mathiesen[1], Thor Einar Andersen[1], Morten Wang Fagerland[1]

[1] *Norwegian School of Sport Science, Oslo, Norway*

In sport science, researchers aim to estimate how cumulative training load affects the risk of sports injury. The strength and direction of training load's effect likely depends on time since exposure, known as a protracted time-lagged effect. Currently employed statistical methods cannot sufficiently estimate such effects. We investigated how to estimate the cumulative, protracted time-lagged effects of training load.
A Norwegian Premier League male football cohort (n participants = 36) with training load data collected longitudinally through a football season (n training load observations = 4 328) was used as basis for simulations. We simulated 7 different scenarios of the effect of training load on injury risk. In each, the probability of injury was a combined result of the magnitude of training load, and the distance in time since the training load exposure. This relationship was simulated with different functions, linear and non-linear, to replicate recent hypotheses of how training load affects injury risk. A Cox regression was run with the simulated injuries as the outcome. Training load was the independent variable and specified with 7 different methods. These methods were either the most frequently used in the training load and injury risk field, recommended in the field, or recommended in environmental epidemiology to estimate similar effects. The simulation was repeated 1 900 times for each of the 7 relationships and the 7 methods. Performance of methods was compared with the Root-Mean-Squared Error (RMSE), Akaike's Information Criterion (AIC), coverage of 95% confidence intervals, and visualizations of the predicted vs. simulated (true) probability of injury.
The distributed lag non-linear model (DLNM) had the lowest RMSE, lowest AIC, and highest coverage across all simulated relationships. The Exponentially Weighted Moving Average (EWMA) had second-to-lowest RMSE and AIC under the assumption that the effect of the exposure, training load, decayed exponentially with time.
DLNM can be used to assess cumulative, protracted time-lagged effects, and is suitable for training load studies in sport science. EWMA may be suitable in studies that have apriori knowledge that the effects decay exponentially over time.

P77

Communicating statistical methods

Estimating surrogacy of time to complete response on PFS in marginal zone lymphoma

Côme Bommier[1], Jérôme Lambert[1]

[1] *INSERM U1153 - ECSTRRA Team, St Louis Hospital, Paris, France*

BACKGROUND: Marginal zone lymphoma (MZL) is an indolent disease for which a surrogate outcome for Progression Free Survival (PFS) is needed. Additionally, the classical meta-analysis methodology cannot be applied due to the few number of randomized trials in this rare disease.
OBJECTIVE: This work uses a slight modification of the single-trial method described by Parast et al. to quantify the surrogacy of time to complete response in 24 months (TTCR-24) for PFS in MZL.
METHODS: Surrogacy is defined as the proportion of treatment effect on the primary outcome explained by the intermediate surrogate outcome information. Moreover, it can further be decomposed between what is added by the surrogate outcome (its incremental value) and what is explained by true outcome information only up to the time of surrogate marker assessment.
Treatment effect is assessed via difference in restricted mean survival time using conditional survival estimates from a landmark Cox proportional hazards model. To handle the fact that the treatment effect has opposite directions on surrogate and true outcome, time to complete response is simply reversed (as landmark time minus time to complete response) in the subgroup of patients that experienced complete response and did not experienced progression or death before landmark time.
RESULTS: In the IELSG-19 phase 3 trial, 401 patients with marginal zone lymphoma were randomized, with a difference in the 8 years restricted mean progression-free survival time of 11 months in favor of experimental arm. At 2 years 82% of RC was observed in experimental arm vs 65% in control arm.
The proportion of treatment effect on 8y-PFS explained by TTCR-24 was 0.95 (95%CI 0.27-1.87), with proportion explained by PFS information up to 24 months equal to 0.75 [0.31,1.27]. The incremental value offered by TTCR-24 over PFS24 was thus 0.20 [-0.23,0.86].
CONCLUSION: Using a simple adaptation of a surrogacy single-trial approach, we showed that TTCR-24 is a valid surrogate marker of PFS at 8 years in marginal zone lymphoma, although its surrogacy is largely explained by the progression events occurring during the first 2 years.

P78

(Semi-)competing risks and causal inference

Estimating Treatment Effects Using Real-World Data When There are Competing Risks

Caroline Chesang[1], Ruth Keogh[1], Linda Sharples[1], Thomas Cowling[1]

[1] *London School of Hygiene and Tropical Medicine, United Kingdom*

Randomized controlled trials (RCTs) are considered to provide the most reliable evidence for establishing evidence for the effects of medical treatments and other interventions. However, the value of evidence on treatment effects from "real-world data" (RWD), such as electronic health records, is increasingly being recognized. Studying treatment effects using RWD requires careful analysis to control for potential sources of bias, including confounding of the treatment-outcome association. Causal inference methods have been developed that enable estimation of treatment effects from RWD, under assumptions.
This work investigates effects of prostate cancer treatments on death from prostate cancer using RWD. In this situation we need to consider competing causes of death, which are particularly common among people with prostate cancer. Defining and quantifying treatment effects when there are competing risks is challenging and needs to account for possible effects of the treatments on other causes of death. Recent work in the causal inference literature has considered this challenge in the context of RCTs, but extensions are required for observational data. This project aims to develop, apply and evaluate statistical methods for the appropriate handling of competing risks, to estimate causal effects of treatments on risk of a specific cause of death using RWD.
In this presentation I will provide an overview of methods for estimating treatment effects using RWD when there are competing risks. I will discuss an application to compare effects of treatment used in prostate cancer and present some preliminary results.

P79

Communicating statistical methods, Meta-analysis

Evaluating the quality of reporting the transitivity assumption in complex networks of interventions

Loukia Spineli[1], Chrysostomos Kalyvas[2], Svenja Seide[3], Katerina Papadimitropoulou[4]

[1] *Hannover Medical School, Germany,* [2] *MSD Europe Inc,* [3] *University Hospital Heidelberg, Germany,* [4] *Leiden University Medical Center, Netherlands*

BACKGROUND: Transitivity assumption is the cornerstone of network meta-analysis (NMA). It implies that important effect modifiers are similar across the observed comparisons in a network of interventions. Violation of this assumption compromises the credibility of the indirect estimates and, by extent, the treatment effects of all possible comparisons in the network. Most of the published empirical studies on the evaluation and reporting quality of the underlying assumptions for NMA have focused on the quality of the indirect comparisons.
OBJECTIVES: The present study is the first to offer extensive empirical evidence on the evaluation and reporting quality of transitivity assumption using a collection of systematic reviews with NMA. The ultimate goal is to elucidate the prevalence of systematic reviews with conclusions of questionable credibility.
STUDY DESIGN AND SETTING: We used a previous collection of 357 systematic reviews with NMA published between January 2011, and March 2017.
RESULTS: Only a handful of reviews reported how they planned to evaluate the transitivity assumption. Two in five reviews explicitly mentioned transitivity assumption and mainly in the discussion section. Most reviews described at least one method to evaluate the transitivity assumption. The most prevalent was the narrative evaluation of trial comparability, followed by sensitivity analysis and statistical evaluation of inconsistency. Almost half of the reviews concluded on the plausibility of transitivity assumption, and hence, the reliability of the treatment effects, followed by the consistency evaluation and intervention hierarchy. Approximately one in five reviews revealed that it was difficult to judge the plausibility of transitivity assumption due to limited available data (e.g., few trials and poor reporting of the included trials). In justifying their conclusions about the credibility of the transitivity assumption, most reviews considered the comparability of the trials, followed by the consistency evaluation. Among the reviews that judged transitivity to be questionable or difficult to judge, only three reviews abstained from performing NMA.
CONCLUSIONS: The prevalence and the methods' quality to assess the transitivity assumption were low overall. We plan to develop a series of recommendations for the comprehensive evaluation of transitivity assumption.

Poster Presentations

P80

Efficient clinical trial designs

Evaluation of an adaptive multi-arm non-randomised sequentially allocated cohort design for phase II

Helen Mossop[1], Michael Grayling[1], Jamal Sipple[2], Martin Thomas[2], Sarah Welsh[3], Grant Stewart[3], James Wason[1]

[1] Newcastle University, Newcastle upon Tyne, United Kingdom, [2] Addenbrooke's Hospital, Cambridge University Hospitals NHS Foundation Trust, Cambridge, United Kingdom, [3] CRUK Cambridge Centre, University of Cambridge, Cambridge, United Kingdom

INTRODUCTION: Efficient phase II trial designs are required to rapidly identify promising agents to take forward into larger trials. Adaptive multi-stage trials are such examples, but their efficiency is reduced if there is a delay in assessing patient response to treatment. Here, we discuss the Multi-Arm Sequential Trial with Efficient Recruitment (MASTER) design, which can provide an efficient and rapid assessment of multiple therapies within one protocol.
METHODS: Motivated by the WIRE trial in renal carcinoma (NCT03741426), we compare three approaches to allocating patients to multiple arms in an adaptive phase II trial: 1) single-arm trials with interim analyses conducted in sequence; 2) a parallel multi-arm multi-stage trial; 3) the MASTER design used in WIRE; in this design recruitment is prioritised to one arm at a time, with a new arm opening to recruitment while interim analyses are undertaken. We conduct a simulation study to compare the time each design takes to evaluate a number of arms, focusing on the setting where recruitment is paused whilst interim analyses are conducted. We investigate how this changes depending on time taken to evaluate the endpoint, the recruitment rate, number of arms, and number of interim analyses.
RESULTS: The parallel multi-arm multi-stage and MASTER design are much more efficient than separate single-arm trials. The average time taken is always shorter with the MASTER design, reducing the time taken to evaluate five treatment arms by around 3-4 months compared to a parallel multi-arm multi-stage design. This advantage is maintained as the recruitment rate and endpoint assessment time increases. The MASTER design provides the greatest gain in efficiency when there is a delay in evaluating the endpoint, or where recruitment rates are moderate to high.
DISCUSSION: We recommend the MASTER design as a promising method to efficiently test multiple potential therapies in non-comparative multi-stage phase II trials. The MASTER design minimises the time recruitment is paused whilst awaiting results of interim analyses and offers additional advantages such as generating complete results on individual arms as the trial progresses. We believe the MASTER design provides a valuable approach to improving the efficiency of early phase trials.

P81

Simulation studies, Other (if none of the topics are applicable)

Evaluation of the Fill-it-up design to combine historical controls and data from RCTs

Stephanie Wied[1], Ralf-Dieter Hilgers[1]

[1] RWTH Aachen University, Germany

CONTEXT: The most appropriate method to assess the effect of an intervention in clinical research is to conduct a randomised controlled trial (RCT). In some areas, for example where small population groups are involved, conducting a RCT is challenging. Historical control data can be included in the new RCT to reduce the sample size of the current trial. There are several approaches to include historical controls in RCTs. However, the approaches usually do not consider their equivalence to justify inclusion.
OBJECTIVES: The aim is to prevent biased estimates of the treatment effect when combining historical controls with control data from a RCT and to reduce the sample size of the RCT. For this purpose, the historical controls should be equivalent to the randomised controls to a sufficient extent to warrant their inclusion.
METHOD: We suggest that historical and randomised controls should only be combined if their expected responses do not differ by a predefined equivalence margin. Our proposed study design is to pause the initially planned study when a specific sample size is reached in order to conduct an equivalence pre-test of historical and randomised controls. If the test confirms equivalence, the controls will be combined for the final superiority hypothesis of the experimental group versus the combined control group. If equivalence cannot be established, the historical controls are not included at all and randomisation of the original trial continues to achieve the required sample size. We compare the presented method with the Bayesian robust meta-analytic-predictive priors[1] (MAP) in the framework of a simulation study with respect to the family-wise error rate and power.
RESULTS: We demonstrate how the total sample size of the initially planned trial can be reduced if historical controls can be combined with randomised controls in the Fill-it-up design. We show under which specific scenarios the family-wise error rate and power of the Fill-it-up design and the MAP approach are kept within acceptable limits.
REFERENCE:
[1] Neuenschwander, B. et al. (2010). "Summarizing Historical Information on Controls in Clinical Trials." Clinical Trials, 7(1), 5–18.

P83

Simulation studies, Other (if none of the topics are applicable)

Evaluation of the performance of propensity score estimators for hierarchical observational data

Andriana Kostouraki[1], Clémence Leyrat[1], Bernard Rachet[1], Aurélien Belot[1]

[1] London School of Hygiene and Tropical Medicine, United Kingdom

When estimating causal effects for observational data, propensity score (PS) methods are used to balance on confounder distributions between treatment groups. These methods include two steps: i) the estimation of the propensity score (i.e., individual probability to receive treatment conditional on postulated confounders), and ii) the use of the propensity score in estimating the causal effect of a treatment on an outcome. Among PS methods, inverse probability weighting (IPW) is the most suitable to account for clustering in the data (such as nesting of patients within hospitals) when the target estimand is the average treatment effect (ATE). Although several estimators have been proposed in this context, their respective performance remains yet unclear under certain scenarios, such as omission of a cluster-level confounder and/or when between-cluster variations in the effects of confounders are observed.
Therefore, we compare via simulations the performance of different estimators when i) a cluster-level covariate is unobserved and/or when ii) between-cluster variation in the effect of a confounder is observed. We assume treatment to be binary and assigned at the individual level with a binary outcome. We analyse the data using PS and/or outcome models with fixed or random effects to account for correlation between patients of the same cluster. We also examine varying effects of an individual-level covariate on the treatment probability by cluster and various sample sizes of the clusters. Results indicate adequate performance of existing estimators when the cluster size is moderate and constant across clusters and the effect of an individual-level covariate on the treatment probability varies between clusters. Challenging remains the case of clusters containing only one treatment level, which entails positivity violations; these are usually excluded and appear when the cluster size is small, and already proposed methods include grouping the clusters according to a predefined criterium; however, even for these methods, the exclusion of clusters including only one treatment level cannot be avoided – this will be further investigated.

P84

(Semi-)competing risks and causal inference, Communicating statistical methods

Exploratory analyses in etiologic research and considerations for assessment of credibility

Kim Luijken[1,2], Olaf Dekkers[1], Frits Rosendaal[1], Rolf Groenwold[1]

[1] Leiden University Medical Center, Netherlands
[2] University Medical Center Utrecht, Netherlands

Exploratory analyses in etiologic research are not always optimally reported. Since these analyses are usually done to generate new research questions, it is tempting to quickly perform a statistical test (or multiple tests) to get a first answer to the problem. However, when such 'quick-test' results are presented in a research article, their interpretation may be ad hoc and unintentionally overconfident. We provide six considerations that should be central to the reporting of exploratory analyses and the discussion about the credibility of exploratory findings. These points focus on the defined research problem, established protocol, cautious assessment of statistical criteria, interpretation of findings, completeness of reporting, and impact of exploratory findings on future etiologic research. The considerations are provided to stimulate a discussion about the preferred handling and reporting of exploratory analyses in etiologic research.

Poster Presentations

P85

Prediction models

External validation of clincial prediction models for transition to psychosis

Laura Bonnett[1], Alexandra Hunt[1]

[1] *University of Liverpool, United States*

BACKGROUND: Psychotic disorders such as schizophrenia are among the world's leading causes of disability. Stratification of individuals meeting high risk mental state criteria according to risk levels presents opportunities for research into underlying mechanisms that may be associated with these risk levels. Various attempts have been made to assess factors associated with increased risk within this population in the form of prediction models. Statistical literature highlights the need to undertake assessment of discrimination and calibration of prediction models in independent data. Therefore, the aim of this study was to assess the external validity of the NAPLS-2 and RAP-I prediction models using the PACE400 dataset.
METHODS: The PACE400 data set consists of individual patient data from 416 individuals meeting at risk mental state criteria who were enrolled in studies conducted at the PACE Clinic with baseline data collected between 1994 and 2006. There were considerable missing data in the PACE dataset so multiple imputation via chained equations was undertaken. The patient characteristics of the development datasets (NAPLS-2 and RAP-I) and the validation dataset (PACE400) were visually inspected to identify differences in scale. Discrimination and calibration were considered for each model to determine external validity in the PACE dataset. The discriminative ability of the model was measured with Harrell's c-statistic and calibration was assessed graphically.
RESULTS: Harrell's c-statistic was 0.55 for the external validation of NAPLS-2 and 0.57 for RAP-I. This suggests a limited to fair ability of the two models to differentiate between those in PACE who developed a psychotic disorder during the study and those who did not. The calibration plots for NAPLS-2 suggest fairly good agreement between the predicted survival probability and the survival probability observed in the PACE400 dataset, while the plots for RAP-I suggest quite poor fitting of the model to the PACE400 dataset with both under- and over-prediction of the survival probability.
CONCLUSIONS: Further work is now required to externally validate the validate NAPLS-2 and RAP-I in alternative external data. If agreement is still sub-optimal it may be appropriate to develop a novel prediction model for use in clinical practice, including external validation of the model.

P86

Prediction models, Other (if none of the topics are applicable)

Factors associated with self-reported asthma among the Canadian population

Barada Mohanty[1], Francis Abayateye[1], Bonnie Janzen[1], Punam Pahwa[1]

[1] *University of Saskatchewan, Saskatoon SK, Canada*

Asthma is one of the most common chronic conditions that adversely impact children and adults' well-being and burden the health care system. Globally, the prevalence of asthma has rapidly increased and doubled in the last decade. The main objective of this study is to determine the prevalence and risk factors associated with asthma in the Canadian population. Data for the present study were obtained from a cross-sectional survey, publicly available in the 2017-2018 (CCHS)-Annual component. Self-reported asthma diagnosed by health professionals was considered the outcome of interest among the Canadian population aged 12 years and above. Exposure factors including demographic, socio-economic and health-related covariates were analyzed by a weighted multivariable logistic regression approach accounting for (i) unequal probability of selection via sampling weights and (ii) design effects (stratification and clustering) via the Taylor Linearization technique. The overall estimated asthma prevalence was 8.05%. Females were more likely to self-report asthma than males (9.12% vs. 6.96%). Obese (ORadj = 1.79; 95% CI: 1.59-2.01) and overweight (ORadj = 1.27; 95% CI: 1.12-1.45) individuals had higher odds of asthma than normal-weight individuals. Similarly, non-smokers (ORadj = 0.78; 95% CI: 0.67-0.89) and ex-smokers (ORadj = 0.82; 95% CI: 0.71-0.95) had lesser odds of asthma than smokers. Individuals with anxiety such as phobia, panic, and obsessive-compulsive disorder (OCD) had significantly higher odds (ORadj = 1.92; 95% CI: 1.73-2.13) of asthma than those with no anxiety-like symptoms. Prevalence of asthma is higher among females than males in the Canadian population; however, the association strongly depends upon other factors such as age, BMI, and smoking status. Factors such as geographical location, immigration status, and household income influence the prevalence of asthma among the Canadian population. As it is hard to get the causal relationship in this study, a longitudinal analysis is recommended.

P87

Communicating statistical methods, Simulation studies

Fitting a three-state disease model to case-cohort data using a weighted likelihood approach

Eddymurphy Akwiwu[1], Thomas Klausch[1], Henriette Jodal[2], Beatriz Carvalho[3], Magnus Loberg[2], Mette Kalager[2], Johannes Berkhof[1], Veerle Coupe[1]

[1] *Department of Epidemiology and Data Science, Amsterdam UMC, Netherlands,* [2] *Clinical Effectiveness Research Group, University of Oslo and Oslo University Hospital, Norway,* [3] *Department of Pathology, Netherlands Cancer Institute, Amsterdam, Netherlands*

INTRODUCTION: Most multi-state models used in describing the natural history process of chronic diseases such as colorectal cancer (CRC) assume that the data is a full-cohort or a simple random sample of the population interest. However, as shown by Akwiwu et al (BMC Med Res Methodol., 2022), such assumptions pose some challenges when fitting these models using data from studies such as CRC surveillance where the proportion of advanced health states is usually low.
OBJECTIVE: We propose a method in fitting a progressive three-state disease model (adenoma-free state [AF] to advanced adenoma [AA] state to CRC state) by weighting the likelihood function using a stratified case-cohort sample in a surveillance setting. We show the gain in efficiency in using a stratified case-cohort sample instead of a simple random sample.
METHODS: Exponential and Weibull distributions were assumed for both transition times. We derived and maximize the weighted likelihood function given a stratified case-cohort sample. Estimates of the asymptotic variance were obtained using a robust sandwich variance-covariance estimator. The performance of the model and the sandwich variance-covariance estimator was assessed using simulated data and the methodology was applied in a Norwegian adenoma stratified case-cohort of size n = 1581 which contains 847 individuals who did not develop CRC after adenoma removal and 734 individuals who developed CRC after adenoma removal.
RESULTS: Estimates of the parameters of the not-directly time distribution from AA to CRC based on the stratified case-cohort sample had negligible biases, good coverage, were more stable in terms of convergence, and showed a gain in efficiency in terms of variance reduction compared to maximum likelihood estimates from a simple random sample by Akwiwu et al. Results from the Norwegian adenoma stratified case-cohort sample (AF: 48.8%, AA: 27.5%, CRC: 23.7%) showed that within 5 and 15 years, about 15.7% (95% CI: 12.4%, 20.0%) and 21.0% (95% CI: 15.3%, 29.1%) of the average aged individuals (66.5 years) will develop CRC, respectively.
CONCLUSIONS: Using stratified case-cohort data in estimating the time distributions in a progressive three-state disease model has been shown to provide more efficient estimates in studies where the proportion of advanced health states is usually low.

P88

Miscellaneous

Flexible modeling of health-related quality of life data accounting for informative dropout

Audrey Winter[1], Benjamin Cuer[2], Thierry Conroy[3], Beata Juzyna[4], Sophie Gourgou[1], Caroline Mollevi[1], Célia Touraine[1]

[1] *ICM, Montpellier, France,* [2] *IRD, Montpellier, France,* [3] *Institut de Cancérologie de Lorraine, France,* [4] *R&D Unicancer, Paris, France*

BACKGROUND: A joint modeling approach is recommended for analysis of longitudinal health-related quality of life (HRQoL) score data in the presence of potentially informative dropouts. A joint model (JM) is composed of two submodels: a model for the time-to-event outcome (e.g., a proportional hazard model) and a model for the longitudinal outcome (e.g., a linear mixed model (LMM)) linked together though shared random effects. The usual LMM modeling the HRQoL outcome is a random intercept and slope model, assuming a linear trajectory over time.
OBJECTIVE: Our aim was to compare, on real data, a modeling based on a linear relationship of the HRQoL score over time, and another, more flexible, using splines.
METHOD: We analyzed 5 dimensions of HRQoL in 267 patients with esophageal cancer enrolled in a randomized clinical trial; global health status, physical functioning, fatigue, pain, and dysphagia. A decrease in questionnaire completion being observed over time, dropout had to be taken into account in order to correctly model the HRQoL; two JM assuming linear or spline-based HRQoL trajectories were applied and compared in terms of interpretation of results, graphical representation, and goodness of fit.
RESULTS: Unlike the JM assuming linear trajectories, the spline-based JM gave a more reliable and precise representation of the HRQoL scores highlighting a nonconstant evolution and variations in the apparent benefit from one arm to the other (with trajectories intersecting once or twice). In our application, it also identified significant arm-by-time interaction effects for almost all dimensions and a significant association effect between the longitudinal outcome and the risk of dropout for one dimension, while these effects were not significant using JMs assuming linear trajectories. Moreover, martingale residuals suggested that the functional form of the HRQoL scores across time in the spline-based JM might be more appropriate.
CONCLUSIONS: The choice of the functional form, when analyzing HRQoL scores over time, turns out to be important, as it affects the validity of the model and the statistical significance. We showed how a linear relationship can lead to wrong or simplistic findings whereas a more flexible modeling gives more reliable and complete results without complicating their interpretation.

Poster Presentations

P89

Prediction models

Forecasting models for healthcare planning suffers from inadequate performance evaluation

Simon Grøntved[1], Mette Jørgine Kirkeby[2], Søren Paaske Johnsen[2], Christina Morh Jensen[1], Jan Brink Valentin[2], Jan Mainz[1]

[1] Region North Psychiatry and Danish Center for Clinical Health Care Research, Aalborg University, Denmark, [2] Danish Center for Clinical Health Care Research, Aalborg University, Denmark

INTRODUCTION: The prediction of future health care capacity needs continuous to be of high demand. Many such models have been developed for disease incidence and prevalence, covering viral diseases such as COVID-19 as well as circulatory and neurological diseases. However, the usefulness and successful implementation of such models in health care planning remains unclear. Thus, following a systematic review and evidence mapping of health care planning, we aimed to describe to which extent such models are validated.
METHODS: All publications retrieved from PubMed and Embase, matching our predefined search strings, were screened for relevance by two independent reviewers using the Rayyan online tool. All publications that could not be excluded by title and abstract were retrieved for full text synthetization.
RESULTS: Of 5415 articles screened 263 were read in full text, of those 105 were deemed eligible for the review. Of the 105 forecasting models 45 was not at all validated, 3 claimed to be validated, but there was no indication of validation in the article. 11 models were only validated by some form of model fit assessment (apparent validation). Temporal validation was performed for 39 models, 4 models were geographically validated, and 3 models were validated with random splits. The most common disease groups are cancer (22 models) and infectious disease (27 models).
DISCUSSION: Most forecasting models for health care capacity needs are not properly validated. Temporal validation is an appropriate method for evaluating performance of a time trend extrapolation, it is also the most commonly used approach. Implementing a model that is not properly validated, can lead to capacity planning decisions that wrongfully allocates resources such that waiting lists in some fields grow, while other fields waste resources on overcapacity.
CONCLUSION: The research on prediction models in health care planning is of limited scope. Meantime, the use of performance evaluation for these models is inadequate. Thus, more research on performance and implementation of forecasting models for health care capacity needs are warranted.

P90

Ageing

Glycemic control prior to cancer incidence and mortality among patients with type 2 diabetes

Jing Haan-Du[1], Klaas H. Groenier[1], Brandy Wauben-Spaetgens[2], Mathilde Jalving[3], Nanne Kleefstra[4], Gijs W.D. Landman[5], Geertruida H. de Bock[1]

[1] Department of Epidemiology, University of Groningen, University Medical Center Groningen, Netherlands, [2] Department of Research and Development, Netherlands Comprehensive Cancer Organization (IKNL), Utrecht, Netherlands, [3] Department of Oncology, University Medical Center Groningen, University of Groningen, Netherlands, [4] Department of Forensic Psychiatry, GGZ Drenthe Mental Health Institute, Assen, Netherlands, [5] Department of Internal Medicine, Gelre Hospital, Apeldoorn, Netherlands

BACKGROUND: An increasing number of patients living with concurrent cancer and type 2 diabetes (T2DM) are followed in primary care, where good quality of glycemic control is feasible and may improve survival. Patients with comorbid cancer and T2DM in the primary care are usually elderly, whose causes of death are usually unknown. This study therefore investigated the role of glycemic control prior to cancer diagnosis on mortality among patients with cancer and T2DM in primary care, accounting for life expectancy relative to the general population.
METHODS: The linkage of the benchmarking database of patients with T2DM (ZODIAC), the Netherlands Cancer Registry, and the Personal Records Database was used (1989–2019, n=71,648). Included were patients diagnosed with incident primary breast, colorectal, or prostate cancer. The target level of glycemic control was defined according to the Dutch guideline. The hazard ratios and 95% CIs for mortality were estimated with multivariable Cox regression model in transformed time. In this model, individual expected survival relative to the general population as well as baseline confounders including age, gender, diabetes duration, history of macrovascular events, body mass index, smoking, socioeconomic status, metformin use, insulin use, lipid-lowering drug use, cancer stage, and baseline year were adjusted.
RESULTS: The numbers of incident cancer patients followed in primary care and the median follow-up years were breast (480, 7.4), colorectal (600, 6.5), and prostate (457, 6.9) respectively. Accounting for the relative survival in the general population, the adjusted hazard ratios for having glycemic control not at target compared with at target among cancer patients with T2DM were breast 0.90 (0.59–1.37), colorectal 1.52 (1.13–2.04), and prostate 1.08 (0.78–1.51) respectively.
CONCLUSIONS: In Dutch primary care for T2DM, a worse glycemic control prior to colorectal cancer diagnosis was associated with 52% increased mortality compared with a good glycemic control, suggesting taking its prognostic value into account. This association was not observed among patients with breast and prostate cancer.

P91

(Semi-)competing risks and causal inference, Missing data

Handling multivariable missing data in causal mediation analysis with a single mediator

Ghazaleh Dashti[1], Katherine Lee[1], Julie A. Simpson[2], John Carlin[1,2], Margarita Moreno-Betancur[1,2]

[2] University of Melbourne, Australia, [1] Murdoch Children's Research Institute, Melbourne, Australia

Longitudinal studies offer an excellent opportunity for causal mediation analysis, but, currently, guidance is lacking on how the common problem of multivariable missing data should be dealt with in these analyses to avoid the well-known risks of bias and loss of precision. Based on a motivating example from the Victorian Adolescent Health Cohort Study, we conducted a simulation study to evaluate the performance of available approaches for handling missing data in one-mediator analyses, where mediation analysis is conducted using either Monte Carlo simulation-based g-computation or inverse probability weighting, both of which require bootstrapping to estimate the confidence interval (CI) with complete data. The missing data methods considered were complete-case analysis, available-case analysis, and five multiple imputation (MI) approaches within the fully conditional specification (FCS) framework, which differed in the interaction terms included in the imputation model, and a "substantive model compatible-FCS" approach. For each MI approach, we obtained the final estimate and CI using two methods: either multiply imputing the data, then within each imputed dataset performing the analysis, including bootstrapping to obtain the standard error, and pooling the results across the imputed datasets using Rubin's rules (known as MI-Boot Rubin); or taking bootstrap samples of the incomplete data, multiply imputing each bootstrap sample and analysing each imputed dataset to obtain the estimate, averaging the estimates across the imputed datasets to obtain the final MI estimate, and using the empirical percentiles of the final MI estimates across the bootstrap samples to obtain the CI (Boot-MI percentile). Our simulations considered two data generation scenarios with moderate and strong interaction terms included in the exposure, mediator, and outcome generation models, and 13 missingness scenarios typically observed in epidemiological studies. Preliminary results suggest that the performance of the missing data methods depends on the missingness scenario and for MI approaches, the strength of the interaction terms in data generation models. We demonstrate the practical value of our findings in an analysis of the extent to which an intervention on common mental disorders (CMD) in young adulthood might reduce the effect of CMD in adolescence on CMD in mid-adulthood.

P92

Machine learning methods for health, Missing data

How do standard imputation strategies impact fairness under clinical presence?

Vincent Jeanselme[1], Maria De-Arteaga[2], Zhe Zhang[3], Brian Tom[1], Jessica Barrett[4]

[1] University of Cambridge, United Kingdom, [2] University of Texas at Austin, United States, [3] University of California, San Diego CA, United States, [4] MRC Biostatistics Unit, University of Cambridge, United Kingdom

Medical modelling aims to improve care for all. Ensuring equal model performance across sensitive groups should therefore be central to statistical models' development and deployment. However, reliance on observational data may encapsulate societal biases against sub-populations, as medical history may have focused on non-representative groups of the population at risk in the development of diagnostic procedures and treatments for a given disease. Moreover, external dimensions such as socio-economic factors or medical expertise may influence the complex interaction between patients and the healthcare system. This phenomenon referred to as clinical presence [1], might influence the process leading to the observed data. Missingness may therefore reflect this process for which assumptions of missing completely at random or at random may be ill-adapted. This raises the question of potential reinforcement of group disparities under different imputation strategies.
This work explores the problem of missingness in medical data and the impact of imputation on group performance. First, we identify three common clinical presence scenarios that result in group-specific missingness in medical data. These are: limited access to quality care, (mis)-informed data collection, and confirmation bias. Then, we make use of simulations to demonstrate how standard imputation strategies can reinforce biases and how no imputation strategy is invariably fairer across clinical presence scenarios.
Our conclusions highlight the importance of carefully considering of how the missingness process of observational data may be occurring, as different imputation strategies may result in group-specific discrepancies in model performance. This analysis encourages practitioners to (1) identify missingness patterns and assumptions made by imputation strategies, (2) perform sensitivity analysis on imputation strategies (echoing recommendation from [2]) and (3) measure fairness metrics at deployment. These are necessary steps towards more equitable care.

REFERENCES:
[1] V. Jeanselme, et al., "Deepjoint: Robust survival modelling under clinical presence shift," arXiv preprint arXiv:2205.13481, 2022.
[2] A. M. Wood, et al., "Are missing outcome data adequately handled? a review of published randomized controlled trials in major medical journals," Clinical trials, vol. 1, no. 4, pp. 368–376, 2004.

Poster Presentations

P93

High dimensional data

Hypothesis testing for CCA given prespecified sparsity levels

Nuria Senar Villadeamigo[1], Mark van de Wiel[1], Koos Zwinderman[1], Michel Hof[1]

[1] *Amsterdam University Medical Center, Netherlands*

High-dimensional data brings multiple challenges to estimation from interpretability to overfitting and singularity concerns. Canonical Correlation Analysis is a popular method that searches for a set of weights maximising the correlation between datasets. When the number of variables exceeds sample size, the probability of selecting highly correlated noise increases and so does the in-sample canonical correlation (CC) estimate. For ill-conditioned problems, the presence of redundant variables is addressed through sparsity methods which threaten stability.
In clinical environments the ability to gather data from each individual has increased, yet the small sample problem persists. Capturing relationships between clinical responses and sets of omics data requires models designed to overcome estimation difficulties arising from high dimensionality.
In this study we focus on testing the hypothesis of no correlation. Since the algorithm for CC is monotonically non-decreasing as a function of nonzero variables, exclusively focusing on this estimate can return less sparse solutions than desired and thus reduce soundness and obfuscate interpretability. Hence, deeming the in-sample CC estimate unreliable for hypothesis testing, we rely on the test estimate. We use k-fold cross-validation to find the optimum CC through a signal of relevant features over a fixed sparsity level. The latter is particularly useful for comparing the test-statistic to the permuted data comprising the null.
Other Sparse CCA models including variable selection methods such as LASSO pay little to no attention to the influence canonical vectors' starting values have on their endpoint. Our algorithm initialises one canonical vector through a grid of random weights or direction vectors based on eigen decomposition.
Since multiple operations have to be iterated, using a fast algorithm, as NIPALS, is essential to constrain computational overload. We use simulations and real data to show the performance of our sparse CCA. We find that for increasingly large matrices fixing the number of nonzeroes brings stability to the shape and size of the canonical vectors compared to other models. We use fixed sparsity levels for hypothesis testing using permutations for the out-of-sample correlation between latent variables to assess the validity of our results and show it maintains the Type I error.

P94

Miscellaneous

Hypothesis testing procedure of sensitivity and PPV for multi-class classification

Kanae Takahashi[1], Kouji Yamamoto[2]

[1] *Hyogo College of Medicine, Japan,* [2] *Yokohama City University, Japan*

Medical tests are important for early detection and disease control in modern medicine. To quantify the performance of the binary medical test, we often use sensitivity, specificity, positive predictive value (PPV), and negative predictive value as performance measures. The sensitivity is the probability that a diseased individual has a positive result. On the other hand, the PPV is the probability of disease when the medical test result is positive. The sensitivity and PPV are also called recall and precision, respectively.
Recall and precision are only applicable to binary classification data. For multi-class classification, two types of aggregate performance measures have been proposed: micro-averaged recall and precision (miR, miP) and macro-averaged recall and precision (maR, maP). The miR and miP pool per-sample classifications across classes and then calculates the overall recall and precision, respectively. Contrarily, the maR and maP calculate an arithmetic mean of the recall and precision for each class, respectively. Although these performance measures were developed in the field of information retrieval, they are gradually being used in the medical field as well.
Most articles report point estimates on recall and precision for multi-class classification without considering uncertainty of the estimates. Moreover, statistical testing procedures for those measures in a paired design have not been proposed yet. Thus, we aim to provide the procedures for comparing the miR, miP, maR, and maP between two medical tests. The performance of the proposed testing procedures is evaluated through simulations.

P95

Meta-analysis

Identifying the Most Effective Components of Lifestyle Interventions for Preventing Type 2 Diabetes

Ellesha Smith[1], Laura Gray[1], Suzanne Freeman[1], Molly Wells[1], Elnaz Saeedi[1], Keith Abrams[2], Daniel March[3], Stephanie Hubbard[1]

[1] *University of Leicester, United Kingdom,* [2] *University of Warwick, United Kingdom,* [3] *University of York, United Kingdom*

BACKGROUND: In many settings, particularly public health, interventions are 'complex' meaning that they are comprised of multiple components. Component Network Meta-Analysis (CNMA) is an evidence synthesis method that was developed to identify combinations of intervention components that are potentially most effective, including combinations that have not been evaluated in previous research.
METHODS: A systematic review identifying evidence of pragmatic interventions for preventing Type 2 Diabetes Mellitus (T2DM) in high risk populations was performed. A component taxonomy was developed with expert opinion that focused on clinically relevant aspects of intervention design, including group or individual setting, modality of delivery and intensity measures. Network Meta-Analysis (NMA) and CNMA models were used to assess the effect of each component for three outcomes: incidence of T2DM, mean change in weight, and mean change in HbA1c.
RESULTS: The systematic review identified 55 studies and nine unique components of interventions: usual care (UC), in person individually (PI), in person as a group (PG), telephone individually (TI), telephone as a group (TG), SMS (SM), app or web-based (AW), email (EM), DVD or TV (DT). Interventions were delivered using between one and five components. The NMAs identified interventions that used PI+PG+TI+SM+DT for intervention delivery as the most effective at reducing weight, HbA1c and incidence of T2DM. The most effective components varied across the three outcomes but there was evidence that PG may improve reduction in weight and HbA1c from baseline to 12 months follow-up.
CONCLUSIONS: This research demonstrated how NMA and CNMA methods can be used to synthesise effectiveness evidence on complex public health interventions and provide more information to policy decision makers for making evidence based recommendations and for informing future research.

P96

Machine learning methods for health, High dimensional data

Immune repertoire features for machine learning-based diagnosis of coeliac disease

Hannah Kockelbergh[1], Shelley Evans[2], Peter Green[1], Andrea Jorgensen[1], Liam Brierley[1], Elizabeth Soilleux[2], Anna Fowler[1]

[1] *University of Liverpool, United Kingdom,* [2] *University of Cambridge, United Kingdom*

In coeliac disease, pathology occurs when T cells respond inappropriately to gluten and self-antigens. This is known as an immune response, where T cells become activated when their receptors bind to specific antigens. T cells somatically rearrange their DNA during development through random selection of gene segments and random addition of nucleotides, leading to a diverse set of immature T-cell receptors capable of binding a wide range of disease-associated peptides. The T-cell repertoire refers to T cells that remain after subsequent selection in the thymus. The CDR3 region is located at the junction of the selected segments and is thought to mediate the bulk of specificity of antigen binding. Currently, coeliac disease diagnosis is subjective and requires prior consumption of gluten and an endoscopy with a biopsy. The T-cell repertoire, when used to train a machine learning classification model, could provide an alternative diagnostic approach for coeliac disease.
Next generation sequencing can be used to capture sequences encoding the CDR3 region of large numbers of T-cell receptors, providing a snapshot of the T-cell repertoire. These CDR3 sequences have variable lengths. Machine learning models trained on immune repertoires tend to suffer from lack of generalisability due to extremely high dimensionality, and limited overlap of CDR3 sequences has been observed between individuals. These factors, together with the structure of the data, incentivise a transformative feature extraction process for immune repertoire datasets.
We capture prospective functionality groups within the immune repertoire by clustering sub-sequences encoded according to physiochemical amino acid properties, such as hydrophobicity. As this feature extraction method is independent of the data, resulting features are likely to be more generalisable. These features offer interpretability in the context of immune receptor functionality and address the issues of high dimensionality and multiple instance format. These qualities contribute to improved performance and utility of immune repertoire classification models for coeliac disease status, allowing more objective diagnosis. We pair this with interpretable machine learning in order to further mechanistic understanding of coeliac disease and discover potential biomarkers.

Poster Presentations

P97

Machine learning methods for health, Missing data

Impact of ALS subtypes on disease progression: A continuous temporal multivariate approach

Juliette Ortholand[1], Igor Koval[1], Sophie Tezenas du Montcel[1], Stanley Durrleman[2], Pierre-François Pradat[3,4]

[1] Brain Institute – INRIA, Paris, France, [2] ICM, France, [3] APHP, Sorbonne University, Paris, France, [4] Northern Ireland Centre for Stratified Medicine, United Kingdom

OBJECTIVES: Amyotrophic lateral sclerosis (ALS) is a rare disease with heterogenous progression rates ranging from slow to rapid progression. Studies aiming to characterize the factors associated with the progression rate have focused on survival but few concerned the functional decline trajectory [1].
METHODS: Using PROACT database (8,569 patients), we select spinal and bulbar patients with at least a baseline and a second follow-up visit. We randomly select spinal patients to get two balanced groups of 1,380 patients.
The following steps were performed: 1) We built a multimodal ALS course map that grasped long-term disease progression in a mixed-effects fashion [2] with Leaspy. We used 6 features: the four subscores of ALSFRS-R, forced vital capacity (FVC) and BMI. 2) We extracted the progression rate and onset age, and the relative progression of each feature of each patient from the parameters of the model. 3) We also computed from the disease course, the conversion age to a progression threshold for each feature: 8 points for ALSFRS-R subscores in reference to FT9 score, 18.5 for the BMI and 2.43 litres for FVC.
Finally, we compared the distributions of the extracted parameters and conversion ages using independent t-test.
RESULTS: We found that ALS starts 2.84 years later for bulbar patients, but progresses 1.36 times faster. We observed, for bulbar patients, that ALSFRS-R fine and gross motor progress 3.7 and 4.4 months later than the spinal patients but ALSFRS-R bulbar progression starts almost 8 months before. Bulbar patients reached endpoint thresholds define above, in average after spinal patients for: ALSFRS-R fine motor (8 points, 49.7 months), gross motor (44 months), bulbar (7.5 months) and respiratory (23.3 months), FVC (24 months).
CONCLUSIONS: We build a modelling framework for describing ALS subtypes effect on functional endpoints from multimodal screening assessments. This model also allows describing and predicting individual progression which can pave the way to discriminate fast and slow progressor for stratification of clinical trial. This methodology could also be applied in clinical trial to compare treated and placebo arms.

REFERENCES:
[1] Grollemund, Vincent, et al. Frontiers in neuroscience 13 (2019): 135.
[2] Schiratti, Jean-Baptiste, et al. The Journal of Machine Learning Research 18.1 (2017): 4840-487

P98

Efficient clinical trial designs, Simulation studies

Impact of duration and number of visits on the power of clinical trials

Emilien Petit[1], Sophie Tezenas du Montcel[1], READISCA

[1] Paris Brain Institute – ICM, France

INTRODUCTION: In the context of rare disease, sample size are limited and clinical trial can turn out to be inconclusive due to lack of power. With a constrained sample size and a given treatment effect, the only leverage to increase power is the duration of the trial and the number of visits. We describe the power of longitudinal clinical trial with the example of SARA progression in spinocerebellar ataxias type 1.
METHODS: SARA progression was estimated using linear mixed models (LMM) on 208 patients included in 3 cohorts (EUROSCA, CRC-SCA and SPATAX). We extracted from our model the intercept, mean slope, variance-covariance matrix of random effects and the residual noise to create simulated longitudinal clinical trial datasets. In each dataset, 2 arms of 30 patients were created and the treatment effect simulated by applying a 50% reduction on the individual slope of patients in the treated arm. Different designs with varying follow-up duration and time between visits were tested. We performed LMMs on the datasets, the treatment effect being tested via the significance of the interaction between time and treatment. 10 000 datasets were simulated for each experience and power was computed as the percentage of significant runs. We studied the impact of the duration, the number of visits within the same duration, the type of the time variable (categorical tested using a Mixed Model Repeated Measures (MMRM) or continuous) and the addition of variability among time interval between visits.
RESULTS: The mean estimated slope of SARA progression was 2.2 per year and the SD of individual random effect on slope was 1.3. Increasing duration drastically increased power for clinical trials from 53% with a 1-year follow-up to 80% at 18 months (visits every 6 months). With the same trial duration, power was increasing with the number of visits. For a 1-year trial with 1-month interval between visits, power reached 80%, gaining 27% compared to the 6 months interval design. Using time as continuous in the LMM compared to MMRM design increased the power of 3% for a 2 years trial. Adding variability among time between visits did not increase power.
CONCLUSION: Power of clinical trials on a linearly progressing outcome can be improved by the trial duration and intermediate visits. However, the benefits-costs balance needs to be considered.

P99

Efficient clinical trial designs

Impact of removing low-information cluster-period cells from a stepped wedge design

Ehsan Rezaei[1], Kelsey Grantham[1], Andrew Forbes[1], Jessica Kasza[1]

[1] Monash University, Melbourne, Australia

BACKGROUND AND AIMS: Standard stepped wedge trials, where clusters switch from the control to the intervention condition in a staggered manner, can be costly and burdensome to both individuals and clusters. Recent work has shown that the amount of information contributed by each cluster in each period differs, with some cluster-periods contributing a relatively small amount of information about the treatment effect. This indicates that 'incomplete' designs, where not all clusters contribute measurements in all periods, may provide sufficient power to detect effects of interest. We investigate whether the same patterns of cluster-period information hold with the iterative removal of low-information cluster-period cells, and how study power correspondingly changes.
METHODS: We remove pairs of cluster-period cells within stepped wedge designs that contribute the least amount of information to the estimation of the treatment effect. We then update the information content of remaining cells, and iterate the removal process until the treatment effect can no longer be estimated. For each design we also compute the power to detect a given effect size, and examine how this changes with progressive removal of cell-pairs. We assume a model for continuous outcomes with a constant cluster-period size, categorical time period effects, and exchangeable and discrete-time decay within-cluster correlation structures.
RESULTS: Cluster-period cells distant from the time of the treatment switches are removed first; removal of these cells does not lead to large reductions in study power. As these cells are iteratively removed, more information is concentrated in the cells near the time of the treatment switch, and in "hot-spots" in the corners of the stepped wedge design. For the exchangeable structure with a high intra-cluster correlation, removing observations from these hot-spots has a marked impact on study power, however this generally does not occur for the discrete time decay structure.
CONCLUSION: Removing cluster-period cells away from the time of the treatment switch may not lead to large reductions in power, implying that certain incomplete designs may be almost as powerful to detect treatment effects as complete stepped wedge designs.

P100

Machine learning methods for health, Prediction models

Implementation of Lightgbm and XGBoost to Survival Data under AFT Model

MD Saifur Mazumder[1], Md. Hasinur Rahaman Khan[1]

[1] University of Dhaka, Bangladesh

CONTEXT: In this modern era, data acquisition has become much easier than before because of modern technologies. As a result, high-dimensional and big data with sparsity is very common everywhere. The sparsity associated with high-dimensional and big censored data become more complex because of the incompleteness due to censoring. In such cases, one of the most important problems of dealing with this type of data is the "Execution Time". It is a very known fact that Extreme Gradient Boosting (XGBoost) is a mighty algorithm that surpassed almost all the existing machine learning algorithms in terms of both performance and functionality. But there is another algorithm called Light Gradient Boosting Machine (Lightgbm) which outperforms XGBoost in terms of execution time.
OBJECTIVE: In this study, we proposed a method to analyze the censored data Lightgbm and also with XGBoost in the Accelerated Failure Time (AFT) model. We identified or predicted the lower risk (observations having the survival time greater than their median survival time) and higher risk (observations having the survival time less than their median survival time) group with both of the algorithms and compare their performance in terms of prediction accuracy and execution time.
METHOD: We used the weighted least square objective function for both algorithms where the weights are calculated from Stute's weights. Several simulation studies have been conducted with three types of distribution (Log-normal, Log-logistic, and Weibull) along with a real-life data set. Both algorithms have been applied with our proposed method and record the test accuracy along with the model training time.
RESULTS: We have used the three types of censoring in the simulation setting (low, medium, and high) and record the result. In almost all simulation settings, Lightgbm has given almost the same accuracy as the XGBoost but it was almost 5-6 times faster than the XGBoost. In the real-life dataset (Diffuse large B-cell lymphoma (DLBCL)), Lightgbm has surpassed the XGBoost by 4% test accuracy.
CONCLUSION: We conclude that it would be advantageous to use Lightgbm with our proposed method to optimize the execution time which will lead to an efficient decision more quickly than the other existing algorithms in survival analysis.

Poster Presentations

P101

Simulation studies

Improving Signal Detection in Small Pharmacovigilance Datasets – A New Pipeline Approach

Hannah Johnson[1], Joan Morris[1]

[1] St George's University of London, United Kingdom

Databases of medications taken in the first trimester of pregnancy and any subsequent birth defects in the fetus are used to detect potential teratogens. Most analyses have examined individual drug-defect associations, compared to overall associations, using methods such as the Proportional Reporting Ratio (PRR), Bayesian Confidence Propagation Neural Network (BCPNN), DuMouchel's Gamma Poisson Shrinker (GPS), and Sequential Probability Ratio Test (SPRT). We compare these methods to a new pipeline approach.
A frequency table of the numbers of pregnancies according to birth defect and medication exposure were simulated for 379 baseline-risk and 21 high-risk (5x baseline risk) drug-defect pairs, for 3 different sample sizes (approximate mean frequency 2.5, 6 and 12 per pair). The power, type 1 error, and positive predictive value of each method were calculated. Simulations were repeated until the Monte-Carlo Standard error was <0.05 for each of these performance measures.
Within the smallest simulated sample size, the new pipeline approach had the greatest power compared to the other methods, while maintaining the type 1 error of the BCPNN. The GPS and SPRT both had lower power, and type 1 error than other methods. Differences in power between the methods reduced with increasing sample size.
This new pipeline approach has increased power in small datasets compared to other commonly used methods, whilst maintaining a conservative type 1 error. This new pipeline method may therefore be of particular use in birth defect datasets which are typically smaller than the large spontaneous reporting datasets that the other methods are tailored toward.

P102

Missing data

Imputation of MNAR variables in an individual participant data meta-analysis

Johanna Munoz[1], Thomas Debray[2], Valentijn De Jong[1]

[1] Utrecht Medical Center, Netherlands, [2] Utrecht University, Netherlands

Missing data is a common problem in medical research, and is often addressed by multiple imputation. While traditional imputation methods allow for valid statistical inference when data are missing at random (MAR), their application is not justified when observations are clustered (e.g. within studies) or when the presence of missingness depends on unobserved information. This situation often occurs when individual participant data (IPD) from multiple studies are combined. Although several imputation methods have been proposed to address individual studies with non-random missing data (MNAR), their applicability and validity in large datasets with clustering remains unclear. Therefore, we propose a new imputation method for multilevel MNAR data. This method is based on the principles of Heckman selection models, and adopts a two-stage meta-analysis approach for the imputation of binary and continuous variables.
To assess the performance of the method and compare it with previously proposed imputation methods, we simulated different scenarios by varying the MNAR specification, the cohort sample size, the number of cohorts and the error distribution. We also illustrated the use of the imputation method in a real multi-district study to estimate the prevalence of malaria in Ugandan children.
Simulation results show that, with the correct specification of the imputation model and exclusion restriction variables, our method generally provides unbiased estimates on parameters that vary across clusters and offers the best coverage among the methods evaluated. However, the method is sensitive to sample size and number of clusters, as well as to deviations from the normality distribution.

P104

Machine learning methods for health, High dimensional data

Investigating use of Random Forest Clustering with multi-omics data to identify novel TNBC clusters

Holly Tovey[1], Jelmar Quist[2], Syed Haider[3], Lucy Kilburn[1], Andrew Tutt[3], Judith Bliss[1], Anita Grigoriadis[2], Maggie Chon U Cheang[1]

[1] Clinical Trials and Statistics Unit, The Institute of Cancer Research, London, United Kingdom, [2] King's College London, United Kingdom, [3] Breast Cancer Research, The Institute of Cancer Research, United Kingdom

INTRODUCTION: Advances in omics technologies has led to an abundance of high-dimensional multi-omics datasets. These come with unique challenges for statisticians due to their highly-correlated, multi-modal nature and require methods developed to integrate different data types. We explored one such method, Random Forest Clustering (RFC) which uses a random forest distance matrix as clustering input, to identify biologically defined subgroups in triple negative breast cancer (TNBC) which may inform treatment options. We assessed impact of parameter choices and compared with consensus clustering (CC).
METHODS: Primary tumours and germline DNA were collected from patients in the Triple Negative Trial and underwent multi-omics profiling: mutation/methylation assays for DNA damage repair (DDR) genes (dichotomous), transcriptional signatures of DDR & immune biology (continuous) and immunohistopathologically assessed tumour infiltrating lymphocytes (discrete). We ran RFC with different parameter (200/500 trees, 100/200 forests, 2/5 features selected in each tree) and variable combinations (average immune score vs individual cell-types, immune vs immune with DDR features). Finally we compared consistency of clusters to those by CC.
RESULTS: Complete data was available for 184 samples. Two dominant clusters were identified with good concordance including 1 enriched for high and 1 for low expression of immune features. Increasing the number of trees & forests did not materially impact clusterships with 175/184 (95%) samples consistently clustered. When increasing the number of variables selected, 168/184 (91%) samples were consistently clustered.
Addition of DDR features did not materially impact immune feature importance or clustership in the 2 main clusters. However, beyond this, DDR features enabled identification of 2 further biologically relevant clusters.
Visual inspection of consensus matrices & cluster assignment tracking plots showed higher consistency in clustering across 100 iterations with RFC compared to CC. Several clusters in CC beyond the main 2 were dominated by dichotomised variables while RFC was not.
CONCLUSIONS: RFC is a reliable method to identify biologically distinct subgroups of TNBC with higher stability than CC in the presence of multi-modal data. Despite low importance, DDR features do inform cluster assignment of TNBC tumours.

P105

Personalized medicine

Is risk of first-episode psychosis mediated by childhood trauma?

Shamika Mohanan[1], Daniel Stahl[1], Marta Di Forti[1], Robin Murray[1], Olesya Ajnakina[1]

[1] King's College London, United Kingdom

Psychosis is a severe mental health syndrome that is characterized by delusions, hallucinations, confused thoughts and paranoia. Psychosis usually presents itself in people suffering from schizophrenia, schizoaffective disorder, schizotypal disorders and delusional disorders. In this project, we assessed the relationship between schizophrenia polygenic risk scores and childhood trauma on the risk to develop psychoses. We analyzed case- control data (n = 1603) collected from the EU-GEI study with 621 cases and 982 controls. European Network of National Schizophrenia Networks Studying Gene-Environment Interactions (EU-GEI) is a large multi-centre study that aims at tracking the development and course of schizophrenia by identifying the clinical, genetic and environmental factors that affect schizophrenia.
We assessed the potential of childhood trauma as a mediator between the exposure- schizophrenia polygenic risk scores and the outcome first-episode psychosis. Two mediator models were developed by analyzing the EUGEI dataset. In one model, childhood trauma was modelled as a continuous latent variable mediator. In the second model, childhood trauma was modelled as a categorical cluster mediator. The cluster mediator variable was developed using finite mixture models. Missing data were handled using full maximum likelihood and multiple imputation. Both models included age, sex, years of education, employment status, relationship status, study site country and the first four principal components from the GWAS study as confounders.
We show that the genetic risk to develop psychoses is partially mediated by childhood trauma. The continuous latent variable mediator model was better at modelling first-episode psychosis case status and had a better model fit. It can be concluded that childhood trauma partially mediates the association between schizophrenia polygenic risk score and risk of first-episode psychosis. In both models, a higher degree of exposure to childhood trauma increases the risk of developing psychosis. As the continuous latent variable mediator fits the data better, it can be suggested that it captures the continuum of childhood trauma better. By modelling childhood trauma as a continuous latent variable, the analysis supports the theory that childhood mediates the genetic risk of psychosis.

Poster Presentations

P106

Meta-analysis

Joint meta-analysis of two diagnostic tests using copula models: application to Alzheimer's disease

Athena Sheppard[1], Rhiannon Owen[1], Sylwia Bujkiewicz[2], Terry Quinn[3]

[1] Swansea University, United Kingdom, [2] University of Leicester, United Kingdom, [3] University of Glasgow, United Kingdom

BACKGROUND: Joint meta-analysis of multiple diagnostic tests answers clinically relevant questions on their comparative accuracy that studying tests in isolation cannot. Modelling the dependencies between two tests evaluated in the same group of patients is vital when assessing the accuracy of different testing strategies, but existing methods are computationally intensive and inflexible to complex dependencies between tests.
Alzheimer's disease dementia is recognised as one of the greatest public health concerns of our time. Changes in multiple biomarkers measured in the cerebrospinal fluid may precede clinically recognisable disease by decades, allowing for earlier diagnosis and the possibility of preventative treatment. The diagnostic accuracy of biomarker tests for Alzheimer's disease must be appropriately assessed.
AIMS: We aimed to develop a meta-analysis model to evaluate the joint accuracy of two diagnostic tests in a Bayesian framework, flexibly capturing the dependence between tests, within studies of the same participants, using a copula function. We aimed to apply these models to a motivating example to assess the accuracy of different biomarker tests for diagnosing Alzheimer's disease.
METHODS: We developed a Bayesian meta-analysis model in R via the package rstan, using a copula function to account for dependencies between two diagnostic tests within studies. We fit the model to diagnostic accuracy data on the comparative performance of biomarker tests for detecting Alzheimer's disease, focussing on modelling summary accuracy measures (sensitivity and specificity) for each test.
RESULTS: Preliminary results suggest that total tau (t-tau) protein was more sensitive (79.8%; 95% CrI: 60.0%, 92.6%) than amyloid beta peptide (Aβ42) (75.6%: 58.8, 89.4), but less specific (t-tau 62.4%: 35.6, 83.2; Aβ42 66.0%: 49.7, 78.2) based on four comparative studies. Hyperphosphorylated tau (p-tau) protein and the ratio of p-tau/Aβ42 also showed promising sensitivity (p-tau: 73.5%: 42.1, 90.8; p-tau/Aβ42: 87.9%: 65.4, 96.6).
CONCLUSIONS: This novel application of copulas to test evaluation appeared to adequately capture dependencies between multiple diagnostic tests within studies. Certain biomarker tests can diagnose Alzheimer's disease with high sensitivity, potentially allowing for earlier detection and intervention in Alzheimer's disease in the future.

P107

Machine learning methods for health

Learning treatment effect in neurodegenerative diseases with a Bayesian mixed-effect model

Pierre-Emmanuel Poulet[1], Bruno Jedynak[2], Stanley Durrleman[1]

[1] INRIA Paris, Paris Brain Institute, France, [2] Portland State University, Portland OR, United States

Disease progression models have successfully been used with longitudinal data from observational cohorts in the field of neurodegenerative diseases. These models recover the natural history of the disease at a population level but are also able to uncover individual heterogeneity which allows for personalized medicine approaches.
However, few have been adapted to cohorts where patients follow a treatment due to the complexity of adding this perturbation to the disease evolution. Since treatment can be either symptomatic or disease modifying, the modelling approaches will differ. Disease modifying implies that history has to be taken into account, while symptomatic can be reduced to a "memoryless" effect.
We use a Bayesian non-linear mixed-effect model as the base disease progression model. Treatment is seen as a time-dependent covariate. Disease modifying treatment effect would imply to also have an history. In this study, we constrict ourselves to modelling only symptomatic treatment effect. This effect is considered as the difference between the observations of patients under the drug and without it. We model this difference as a function of several variables, including the dose and the current disease stage of the patient. We learn this function as a regression using radial basis functions.
We applied our approach to Parkinson's disease with the Parkinson's Progression Markers Initiative cohort. We consider the dopaminergic treatment as a symptomatic effect on the disease outcomes. Our goal was to learn the differences between "ON" and "OFF" states (under or not under drug effect) and to understand when the treatment reaches its maximum efficacy. Results show that dopaminergic treatment reduces motor impairment best when the patient was at a medium-advanced stage of the disease (MDS-UPDRS part 3 motor examination around 40). Higher dosage does not imply better motor score improvement. Our experiments suggest that the best variable to explain the efficiency of the treatment is the current disease stage of the patient.

P108

Meta-analysis

Likelihood-based inference in control risk regression with study-specific covariates

Phuc Tran[1], Annamaria Guolo[1]

[1] University of Padova, Italy

Detecting heterogeneity among studies about the same issue of interest is one of the main goals in meta-analysis. When the interest is the evaluation of the effectiveness of a treatment, the between study heterogeneity can be explained using control risk regression, a meta-analysis model that includes a measure of risk for subjects in the control condition. The measures of risk for the treatment group and the control group are affected by errors, as they are summary information from each study included in the meta-analysis. Properly accounting for measurement error is necessary for inference to be reliable and the literature has recently focused on some parametric and non-parametric solutions.
Nevertheless, an appropriate explanation of the between-study heterogeneity could require the inclusion of additional study specific characteristics, or covariates. When the information is available at the aggregated level, and not at the individual level, the measurement error problem can extend to the additional covariates, leading to methodological and computational issues. We explore the effects of errors affecting covariates in control risk regression on the results from a likelihood-type inference, a topic which received less attention in the literature. When aggregated information from subgroups within each study is available, we derive appropriate approximations of the within-study variance/covariance components associated to the distribution of the error-affected covariate. When the information is not available, the standard likelihood approach cannot be applied, and a pseudo-likelihood solution is proposed, that simplifies the likelihood function under a working independence assumption between the involved variables.
The proposed approaches, and the pseudo-likelihood solution in particular, have a satisfactory performance, after evaluation through extensive simulation studies under different scenarios, including small sample size and different baseline risk distributions. Methods are also applied on a real dataset of serum cholesterol reduction.

P109

Ageing

Long-term trends in height and body mass index in Japan

Ikuko Funatogawa[1]

[1] The Institute of Statistical Mathematics, Tokyo, Japan

INTRODUCTION: COVID-19 became a global pandemic, and obesity is a recognized risk factor for severe COVID-19. The prevalence of obesity is high in many countries, and obesity is currently a worldwide concern. However, Japan maintains a relatively low prevalence of obesity. In our previous studies, we showed changes in body mass index (BMI) by birth cohort in Japanese adults and Japanese women aged 1 to 25 years based on a repeated cross-sectional nationwide survey (Funatogawa et al., 2008, BMJ; Funatogawa et al., 2009, Int J Epidemiol). More than 10 years have passed since the previous studies, and here we show the long-term trends in Japan with adding data in new survey years.
METHODS: We analyzed data from the National Nutrition Survey, Japan (NNS-J) for subjects aged 1 to 25 years, 20–29 to 60-69 years, and 65–69 to 80-84 years. Height, weight and BMI during 1950–2019 were estimated according to the age groups and birth cohorts.
RESULTS: The height was higher in newer birth cohorts for both men and women, but recently there is not much difference across birth cohorts, and a slight decrease is seen in young ages. BMI in all adult age groups increased from 1890 (1895-1904) to 1990 cohorts for men. BMI in women increased until 1930 or 1940 cohorts, but later decreased until 1970 or 1980 cohorts. In both men and women, BMI in 1980 cohort is similar to that in 1970 cohort, BMI in 1990 cohort (at ages 20-29) is higher than that in 1980 cohort, and BMI in 2000 cohort (up to age 15) is lower than that in 1990 cohort.
CONCLUSION AND DISCUSSION: In Japanese men, newer birth cohorts have higher BMI in adulthood. The average BMI at ages 20-29 is about 23 for men and about 21 for women in the recent cohort. In Japan, not only obesity but also low birthweight and thinness of young women are also health concerns, and the guideline for weight gain during pregnancy was changed in 2021 to reduce the low birthweight. Since BMI has increased in all birth cohorts from ages 20-29 to 50-59 in this study, careful attention must be paid to the possible increase in obesity several decades later.

Poster Presentations

P110 Machine learning methods for health, Prediction models

Machine learning at the service of survival analysis: predictions using time-to-event decomposition

Lubomir Stepanek[1], Filip Habarta[2], Ivana Mala[2], Lubos Marek[2], Ladislav Stepanek[3]

[1] *First Faculty of Medicine, Charles University, Prague, Czech Republic*
[2] *Faculty of Informatics and Statistics, Prague University of Economics and Business, Czech Republic*
[3] *Faculty of Medicine and Dentistry, Palacký University Olomouc, Czech Republic*

Predictions of lengths of the event of interest-free intervals, based on multiple covariates values, are routine in survival analysis and commonly used in biostatistics. There is a toolbox of methods crunching the task, particularly the Cox proportional hazard model and its numerous variants. However, all the methods have limitations, and their application is determined by meeting statistical assumptions.
In this work, we address the established methods' statistical limitations and retake the task of the event of interest-free intervals' lengths predictions differently, enabling us to apply machine-learning techniques, having a minimum of assumptions, on a decomposed time-to-event variable, describing the event of interest time distribution. The proposed decomposition allows us to understand the time-to-event variable as two new variables. While the first component describes whether the event of interest occurred or not in a given individual and if so, the second component then depicts when exactly the event occurred. Thus, whereas the first component is a binary one, determining if the event of interest did happen or not, the second one is continuous, since measuring the time to event of interest, or censoring, if any. Consequently, we refine the event of interest occurrence prediction as a supervised classification task and the time-to-event prediction as a supervised regression task. So, we might construct predictions of event occurrences and time-to-event intervals by applying regression and classification algorithms such as multivariate or multinomial regression, support vector machines, regression and classification trees, random forests, naïve Bayes classifiers, and neural networks.
Using the COVID-19 dataset of many explanatory covariates and a response variable as a time to COVID-19 antibody blood level decrease below a laboratory cut-off, we built the machine-learning regression and classification models and predicted the time to antibody decrease below the cut-off. Moreover, we also estimated the time-to-event intervals using the Cox proportional hazard model.
Comparing both approaches and getting similar performance metrics' values, the proposed machine-learning predictions based on time-to-event variable decomposition seem to be a valid alternative to established Cox regression, minimizing required statistical assumptions.

P111 Machine learning methods for health, Prediction models

Machine learning methodologies for modelling of time-to-event endpoints in Prostate Cancer

Autumn O'Donnell[1], Shirin Moghaddam[2], Eric Wolsztynski[1],

[1] *University College Cork, Ireland,* [2] *University of Limerick, Ireland*

Prostate cancer is among the most common cancers for men globally, accounting for 13% of cancer diagnoses in the male population each year. Radical prostatectomy (RP) fails in 20-40% of patients, who subsequently develop biochemical recurrence (BCR). This study focuses on machine learning methodologies for prediction of major prognostic endpoints, including in continuous time such as time-to-BCR prediction. An overarching aim of this work is the inclusion of genetic attributes alongside routine clinical information.
Analyses were carried out on a clinical cohort of 198 patients with BCR-free time information, 140 of which also had mRNA data. Predictive models used in the benchmarks included tree-based methods and other learning strategies such as boosting and regularization applied to Cox proportional hazards modelling. Other steps of the machine learning pipelines, including the choice of prefiltering technique and the relative contribution of candidate feature for prediction of BCR-free survival from both pre-RP and post-RP perspectives, were also assessed via comparative analyses. Potential interpretation of these features was examined in the context of such multivariate models.
Preliminary cross-validated results indicated that inclusion of mRNA information yielded increased prognostic performance with a 5% increase in C-index (95% Confidence Interval: (1%, 11%)), different performances across prognostic modelling strategies, and at various clinical time points (e.g., post-operative prediction outperformed pre-operative performance (C-index: 0.80 vs 0.73; 95% CI: (-0.14, -0.01)).
These levels of performance were also considered against routine predictive nomograms to assess their clinical potential for integration in routine practice.

P112 Prediction models, Personalized medicine

Machine learning performs better in individual dynamic prediction with non-proportional hazards

Baoyi Huang[1], Chengfeng Zhang[1], Yawen Hou[2], Zheng Chen[1]

[1] *Southern Medical University, Guangzhou, China,* [2] *Jinan University, Tripoli, Lebanon*

The existing static models at baseline may have limitations when handling non-proportional hazards covariates. With longitudinally changing in patients' condition, the values and effects of covariates may change dynamically that did not satisfy proportional hazards assumption. Therefore, establishing a dynamic prediction model to utilize such updating information and make dynamic individual prediction is necessary. We introduced a dynamic Cox model and a dynamic machine learning model both basing landmarking approach, making their prediction and comparison. Two examples with time dependent covariates and covariates with time-varying effects were considered to be analyzed. The prediction performances of two models were assessed by C-index and Brier score, and how to make individual dynamic prediction will be shown. We found out that the dynamic machine learning model performed better than the dynamic Cox model which might predict survival probabilities slightly high. In conclusion, dynamic prediction models can analysis these non-proportional hazards covariates and make prediction in different time points both of which are limitations of static models. The use of dynamic machine learning model make predictions more accurately and provide evidence for clinical decision making that serving as a better tool for individual dynamic prediction.

P113 Simulation studies, Prediction models

Mathematical-Model Analyses of the Effects of Rapid ART on HIV and AIDS

Kazumi Omata[1], Toshibumi Taniguchi[2], Mayumi Imahashi[3], Tatsuya Noda[4], Yoshimi Kikuchi[1], Shin-Ichi Oka[1], Hiroaki Mitsuya[1]

[1] *National Center for Global Health and Medicine, Tokyo, Japan,* [2] *Chiba University Hospital, Japan,* [3] *Nagoya Medical Center, Japan,* [4] *Nara Medical University, Japan*

Rapid antiretroviral therapy (ART) is expected to contribute not only to the treatment of people living with HIV, but also to the reduction of new infections. We estimated the number of future infections using two mathematical methods. In the first method (Method I), the number of new infected people is estimated based on the back- calculation method and extrapolation. The second method (Method II) uses the compartment model which employs four compartments: susceptible, infectious, AIDS, and treated. The number of new HIV infections and AIDS cases reported by the AIDS Prevention Information Network in Japan, and clinical information at the National Center for Global Health and Medicine were used as input values. It was assumed that Rapid ART will be actively launched in 2023, and the rate of detection of infected people by testing will remain the same. When the ratio of suppressing the viral load by Rapid ART was set to 81% in Method I or the average diagnosis-to-treatment period was set to 24 months in Method II, it was calculated that the number of infected people after 2025 was almost flat. Here, 81% is a numerical value assuming the medical treatment rate of 90% and the virus control rate 90%, that is, the second and third ones of the UNAIDS 90-90-90 targets. When the ratio was greater than 90% or the period was set to 6 months, the number of new infected people decreased quickly with the start of Rapid ART, and a gradual phenomenon was observed. Rapid ART has been shown to have a significant impact on the decrease in the number of infected people, but even if the medical treatment rate and virus control rate can be achieved, if the infection rate by the test remains as it is, the number of infected people may not be expected to decrease significantly, indicating that it is necessary to improve the test system. And if we expect to end the spread of the infection, we may need to look at the problem on a decades-long scale.

Poster Presentations

P114

Prediction models, Personalized medicine

Measuring the performance of prediction models to personalize treatment choice

Orestis Efthimiou[1], Jeroen Hoogland[2], Thomas Debray[3], Michael Seo[1], Toshiaki A Furukawa[4], Matthias Egger[1], Ian White[5]

[1] Institute of Social and Preventive Medicine (ISPM), University of Bern, Switzerland, [2] Julius Center, University Medical Center Utrecht, Utrecht University, Netherlands, [3] Utrecht University, Netherlands, [4] Departments of Health Promotion and Human Behavior and of Clinical Epidemiology, Kyoto University, Japan, [5] MRC Clinical Trials Unit at UCL, United Kingdom

Personalized decision-making has attracted much interest lately. When data are available from individual patients receiving either a treatment or a control intervention, various statistical and machine learning methods can be used to develop prediction models to personalize treatment choice. Such models aim to predict future outcomes of patients under the two interventions, and can thus provide individualized estimates of treatment benefit and help personalize treatment choice. However, relatively little attention has been given to methods for measuring the performance of predictions of personalized treatment effects. In this talk, we propose a range of measures that can be used to this aim. We start by defining two dimensions of model accuracy when aiming at treatment effects: discrimination for benefit, related to the ability of a model to differentiate between patients who will benefit more from patients who will benefit less from treatment, and calibration for benefit, i.e. the ability of a model to estimate the magnitude of treatment effects. We then propose an additional concept, decision accuracy. This amalgamates these two dimensions and describes the ability of a model to identify patients who should receive treatment with respect to a threshold for benefit, i.e. a threshold above which treating patients is deemed worthwhile. Subsequently, we propose a series of estimands related to these three dimensions and discuss estimating procedures, focusing on the case of randomized data. Our methods are applicable for continuous or binary outcomes, and for any type of prediction model. We implement all methods in the R package predcompare. We illustrate all methods using simulated data and a real dataset from a large randomized trial in depression. Our results suggest that the proposed measures can be potentially useful in evaluating and comparing the performance of multiple competing models with respect to their ability to predict individualized treatment benefit. We conclude that decision accuracy is an easily computed measure, which can help in determining the value of personalised clinical decision-making.

P115

Simulation studies, Meta-analysis

Methods for modelling the multi-state natural history of rare diseases using disparate IPD sources.

Jonathan Broomfield[1], Suzanne Freeman[1], Keith Abrams[2], Nick Latimer[3], Mark Rutherford[1], Michael Crowther[4,5]

[1] University of Leicester, United Kingdom, [2] University of Warwick, United Kingdom, [3] School of Health and Related Research (ScHARR), University of Sheffield, United Kingdom, [4] Red Door Analytics, Stockholm, Sweden, [5] Karolinska Institute, Stockholm, Sweden

BACKGROUND: Multi-state survival models are often used to represent the progression of a disease across a patient's lifetime. These models are termed natural history models, and can be used in a health technology assessment to represent the current standard of care against which a novel treatment can be compared. The construction of these models for rare diseases is problematic since evidence sources are typically much sparser and more heterogeneous.
OBJECTIVE: This simulation study investigated different one-stage and two-stage approaches to meta-analysing individual patient data (IPD) in a multi-state survival setting when the number of studies meta-analysed is small, and the studies vary in size. The objective was to investigate how accurately simpler methods (that make more restrictive assumptions) can model the data, and when more computationally intensive models are less likely to converge. The study also investigated how much larger studies can dominate the analysis over smaller studies, or whether the true population-wide estimates of disease progression can still be recovered.
METHODS: Multi-state IPD were simulated under a variety of data-generating mechanisms to represent multiple studies with (potentially) different baseline Weibull transition-specific hazard functions. Parameter values were informed using real-world evidence on Duchenne Muscular Dystrophy (DMD). Hazard functions using one-stage frailty and two-stage stratified models (the second stage being a multivariate meta-analysis of transition- and study-specific hazard parameters) were estimated, and compared to a "base case" model that did not account for study heterogeneity. The bias, coverage and empirical standard errors of transition probabilities to, and lengths of stay in, each state at different points in follow-up (as well as model convergence) were used to assess model performance. The Stata commands survsim and merlin were used to simulate data and fit models. A real-world application of methods to DMD was also conducted.
RESULTS AND CONCLUSIONS: Frailty models are difficult to estimate with small samples, particularly when the number of studies is small, even when they are the "correct" model. Studies that have missing data on some transitions can also restrict model convergence. Stratified models appear to offer a compromise between model accuracy and robustness.

P117

Simulation studies, Missing data

Missing outcomes in prognostic prediction models: A simulation study

Mauricio Moreira Soares[1], Erlend Fossen[1], Arnoldo Frigessi[1], Marissa LeBlanc[1]

[1] University of Oslo, Norway

A common problem for prognostic prediction model studies is missing outcomes. Often machine learning studies avoid using explicit imputation methods in favour of complete case analysis. This implicitly assumes missingness completely at random, and even if this assumption holds, complete case analysis is inefficient. If the missingness is not completely at random, it is well known that the complete case set is a selected, non-representative sample and will result in biased estimates of association and predictive performance. The aim of this study is to evaluate the role of different methods for handling missing data in prediction models with missing outcomes. We will assess how handling the missing data (e.g. multiple imputation) prior to data split into training and testing sets can impact the models performance measures. For this purpose we use simulated data generated with a decision tree model based on statistics available from the Head and Neck 5000 (H&N5000) study. We used selected clinical and demographic variables and a binary outcome. After generating missing values (MCAR; MAR; MNAR) we calculate the prediction performance with AUROC for logistic regression models and explore the following situations: full dataset (AUROC-0), imputed cases with the whole data (AUROC-1) and performing imputation after splitting (AUROC-2). Then we compare the AUROC-n for each imputation strategy (n=1, 2) with the full dataset AUROC-0 to determine the impact of imputation: an increase in model performance indicates the introduction of non-independence during imputation. We repeat this analysis also for empirical data from the H&N5000 study. This study provides insights about the effects of different methods for handling missing data on prediction performance.

P119

Prediction models, Personalized medicine

Mixed-effects location-scale models for within-individual variability in cystic fibrosis

Marco Palma[1], Jessica Barrett[1], Graciela Muniz Terrera[2], Ruth Keogh[3], Angela Wood[4]

[1] MRC Biostatistics Unit, University of Cambridge, United Kingdom, [2] Centre for Clinical Brain Sciences, University of Edinburgh, United Kingdom, [3] The London School of Hygiene & Tropical Medicine, United Kingdom, [4] Cardiovascular Epidemiology Unit, University of Cambridge, United Kingdom

The longitudinal trajectories of lung capacity are the main outcome of interest to predict life expectancy in cystic fibrosis patients. While the mean trajectories over the lifespan have been well studies in the literature, the fluctuations of the longitudinal measurements over time are either neglected or treated as a fixed/static quantity by using some basic summary statistics (e.g. standard deviation or coefficient of variation). Within-individual variability (WIV) in lung function could play a relevant role in understanding disease progression and detecting sudden lung functions drops which require intensive treatment.
We use a mixed-effects location-scale model (MELSM), an extension of linear mixed models where both the mean and the within-individual variance depend on covariates and random effects. The model allows for correlation between the patient-specific random effects in the location and in the scale submodels. A Bayesian multilevel approach is used to assess variability for the individual random effects. The model is applied on the latest version of the CF Epinet dataset, which includes annual measurements of lung capacity for 13899 UK patients. The analysis highlights some age-sex interactions in the WIV over the lifespan at the population level. Furthermore, it shows that both the patient-specific random effects in the location and scale submodel are not negligible and mildly correlated. The results open the way for an improved prediction of risk adverse outcomes via joint models.

Poster Presentations

P120

(Semi-)competing risks and causal inference

Model selection for multi-state models in medical research

Kaya Miah[1], Axel Benner[1], Annette Kopp-Schneider[1]

[1] *Division of Biostatistics, German Cancer Research Center (DKFZ), Heidelberg, Germany*

For modeling the holistic disease progression process as well as considering the complexity of modern therapy concepts in hematological oncology, multi-state models reveal a suitable methodological approach. Especially in the context of complex event time structures along with increasing covariate information, such a differentiated analysis is of great interest.
Thus, statistical methods for model selection within complex multi-state models are essential to more accurately capture pathogenic disease processes as well as underlying etiologies. An important goal is to select sparse models based on higher-dimensional molecular data with minimal loss of predictive ability.
Common methods for model selection incorporate penalized regression models or boosting. Especially in higher dimensions, statistical learning algorithms reveal powerful techniques with respect to model selection. In the context of stratified Cox regression formulation of multi-state models, a promising idea for reducing model complexity is to combine homogenous covariate effects for distinct transitions based on reparametrization of the models along with data-driven variable selection by regularization methods.
This raises the following challenges: First, multiple transitions have to be adjusted for consecutive treatment phases within the multi-state model. Furthermore, the number of submodels with decreasing number of observations increases with the complexity of the model during the course. Finally, model selection approaches have to be implemented for multi-state settings in higher dimensions.
Thus, the extension of model selection procedures for complex multi-state models utilizing higher-dimensional data provides a promising tool for a more precise understanding and interpretation of individual disease progression, specific oncological entities along with their therapeutic concepts as well as improved personalized prognoses.

P122

Communicating statistical methods

Modelling disease transitions in multimorbidity via multistate models

Sida Chen[1], Jessica Barrett[1], Christopher Jackson[1], Tom Marshall[2], Krish Nirantharakumar[2], Francesca Crowe[2]

[1] *MRC Biostatistics Unit, University of Cambridge, United Kingdom,* [2] *University of Birmingham, United Kingdom*

Multimorbidity, the co-existence of two or more chronic conditions in an individual, is becoming increasingly prevalent in ageing societies and imposes significant challenges to individuals and public health. It is important to understand how diseases accumulate over time and the effects of potentially relevant risk factors on comorbidity onset and progression. To date, however, research has mainly described cross-sectional patterns of multimorbidity. Here we proposed the use of a spline-based parametric multistate modelling (MSM) framework for modelling and predicting the temporal evolution of diagnosed chronic conditions. The model is highly flexible and permits a parallel estimation framework under mild assumptions. We used the MSM to analyse a large UK electronic health record dataset (Clinical Practice Research Datalink) involving a total of more than 13 million eligible patients, where we examined the effects of several demographic and socioeconomic risk factors on the progression of multimorbidity defined from five chronic conditions, namely the cardiovascular disease, type-2 diabetes, chronic kidney disease, mental health and heart failure. Our results highlight the complex influences of these risk factors on the transitions between different comorbidity states and bring new insights into how these effects are modified through comorbidities.

P123

Miscellaneous

Modelling non-linear time-varying intervention effects on recurrent events

Xiangmei Ma[1], Yin Bun Cheung[1], KF Lam[2], Chee Fu Yung[3], Paul Milligan[4]

[1] *Duke-NUS Medical School, Singapore,* [2] *University of Hong Kong,* [3] *KK Women's and Children's Hospital, Singapore,* [4] *London School of Hygiene & Tropical Medicine, United Kingdom*

A non-linear function for modelling time-varying intervention effects was proposed in a study of infectious disease prevention. It was a flexible yet interpretable four-parameter monotonic function of log hazard ratio. The monotonic function assumed protective efficacy to change from zero to peak short-term level practically immediately upon receipt of the intervention. For slower acting interventions, we consider an alternative four-parameter non-monotonic function, which allows the initial impact to take time to build up. In addition to two parameters (a, d) that represent the initial and long-term impact caused by the intervention, two other parameters (phi1, phi2) represent the damping rate and strength of restoring force in an oscillatory movement, which describe the shape of the non-linear pattern. The cumulative effects of multiple intervention doses over time can be captured by a sum of a series of the function, with each component of the series capturing the time-varying effects of the latest dose. For illustration, we use data from a trial of four doses of seasonal malaria chemoprevention given over a duration of nine months, where the (recurrent) outcome events were clinical malaria episodes. We apply the Andersen-Gill model with a sum of series of either the monotonic or non-monotonic time-varying effect functions and compare the results with cubic B-spline models with various complexity and step functions that allow for high degree of flexibility but less interpretability. Both AIC and BIC indicate that the model with the non-monotonic function has the best fit. It is estimated that it took about 10 days after dosing to reach the peak efficacy level, which dropped by half in about 31 days. Simulations show the parameter estimators are asymptotically unbiased and the confidence intervals achieve target level of coverage probability as either the number of subjects or number of doses of intervention increase.

P124

Communicating statistical methods

Modelling the impact of SARS-CoV-2 vaccination in a cohort of patients hospitalized for COVID-19

Natasa Kejzar[1], Dasa Stupica[2]

[1] *Faculty of Medicine, University of Ljubljana, Slovenia,* [2] *Department of Infectious Diseases, University Medical Centre Ljubljana, Slovenia*

The data about the cohort of adult patients (≥18 years) hospitalised in autumn 2021 and spring 2022 were collected to evaluate the association between primary vaccination, SARS-CoV-2 variant (Delta/Omicron) and progression to critically severe disease (mechanical ventilation or death). Not all the patients achieved an endpoint during observation period. Descriptive group comparison showed that fully vaccinated patients were older, more often immunocompromised, and had higher Charlson comorbidity index scores.
We analysed the outcome (critically severe disease) using logistic regression statistical model (adjusted for selected covariates) and complemented it with the analysis of propensity-score matched sample. Due to missing values, we supplemented the analysis with (multi-state) time-to-event analyses models.
We present the models and highlight some issues that we have encountered when trying to interpret the results.

Poster Presentations

P125

Efficient clinical trial designs

Moderation analysis in an RCT with repeated measurements: Constrained longitudinal analysis.

Stian Lydersen[1]

[1] *Norwegian University of Science and Technology, Trondheim, Norway*

In an RCT with measurements of the outcome variable at baseline and one or more later time points, a linear mixed effects regression model can be suited. In general, including a strong predictor of the outcome in an RCT as a covariate can improve substantially the precision of the estimated treatment effect. The baseline value of the outcome variable is usually a strong predictor of the outcome at a later time point. In an RCT with repeated measurements, this can be handled by including the baseline value as part of the outcome variable in a linear mixed model, and without including a main effect of the treatment variable at baseline, rather than including the baseline as a covariate. Many researchers are not aware of this. For example, if the outcome variable is measured at time 0 and two later time points, it is modelled as $Yij = \beta 0 + \beta 1t1 + \beta 2t2 + \beta 3t1x + \beta 4t2x + Bj + \varepsilon ij$, where ti is an indicator variable equal to 1 at time i, x is an indicator for the treatment group, Bj is the random effect of participant j, and εij is the residual term. We assume that the baseline measurement was done before, or blinded for, randomization, so we can assume that there are no systematic differences between the groups at baseline. This way of handling the baseline value is called constrained longitudinal analysis by Coffman &al, BMJ Open 2016, e013096.
In an RCT, it may be relevant to investigate whether certain subgroups of patients have more benefit of the treatment than others. We have generalized the concept of constrained longitudinal analysis to moderation analysis, and we will illustrate this approach in a secondary analysis of a large RCT.

P126

Miscellaneous

Monitoring time to event in medical registry data using CUSUMs based on excess hazard models

Jan Terje Kvaløy[1], Jimmy Huy Tran[1]

[1] *University of Stavanger, Norway*

In health registries, for instance cancer registries, patient are routinely registered at diagnosis, and outcome data like survival times are often added later. Based on such data an aspect of interest could be to monitor whether the distribution of the time to an outcome of interest changes over time – for instance if the survival times of cancer patients changes over time, while adjusting for known risk-factors. Such monitoring could be of interest both for real-time monitoring of incoming data, and for retrospective analyses to pinpoint when in time important changes took place.
A common challenge in monitoring survival times based on such registry data is that time to death, but not necessarily cause of death is registered. To quantify the burden of disease in such cases, excess hazard methods can be used. With excess hazard models the total hazard is modelled as the population hazard plus the excess hazard due to the disease. The population hazard is found from national life tables.
We propose a CUSUM procedure for monitoring for changes in the time to event distribution in such cases where use of excess hazard models is relevant. The procedure is based on a survival loglikelihood ratio, and extends previously suggested methods for monitoring of time to event to the excess hazard setting. The procedure takes into account changes in the population risk over time, as well as changes in the excess hazard which is explained by observed covariates. Properties, challenges and an application to cancer registry data will be presented.

P127

Medical device clinical studies, Missing data

Multiple imputation by of missing glucose monitor data by Brownian Bridge stochastic interpolation

Colin Everett[1], Rebecca Walwyn[1], Julia Brown[1], Leonid Bogachev[2], Jeanne Houwing-Duistermaat[2], Deborah Stocken[1]

[1] *Clinical Trials Research Unit, University of Leeds, United Kingdom,* [2] *School of Mathematics, University of Leeds, United Kingdom*

INTRODUCTION: Continuous glucose monitoring devices record high-frequency glucose levels over many days. If the sensor is detached from the wearer, or storage memory is filled, then data may become missing. Current recommendations for analysis of partially-complete stream data make strong and implausible assumptions around the distribution and variability of unobserved values. It is therefore desirable to explore an approach to handling such missing data while making more relaxed assumptions.
METHODS: Missing intermediate values were multiply imputed by linear interpolation with a random element, step-by-step until the data was observed again. The conditional mean was updated at each step and, like the random variates could depend on participant characteristics and time of day. To illustrate its use, simulated glucose data traces were generated, having intra-day drift and volatility patterns over two weeks. Incomplete datasets for each of five different missing data patterns were then derived from complete data by deleting data under patterns with differing dependence on times of day, patient characteristics and true values. Three glucose control measures were derived for each patient in each dataset after applying five methods of accounting for the unobserved data: i) extrapolating from observed data, ii) simple linear interpolation of missing data, iii) last observation carried forward and iv) Brownian Bridge multiple imputation with and v) without accounting for participant trend patterns. Bias, empirical standard error and mean squared error of estimated mean glucose, time in “normal” range and coefficient of variation of glucose for the single imputation and stochastic Brownian Bridge multiple imputation (after 10 imputations) methods were calculated.
RESULTS: All methods exhibited bias when analysing data under all missing data patterns. Preliminary results suggest that Brownian Bridge interpolation often underestimates mean glucose, with bias comparable to simpler methods. It appeared to account for uncertainty due to unknown data. Simply extrapolating from the observed data led to large bias in estimated TIR when data was missing not at random. Comprehensive results will be presented.
CONCLUSIONS: It is feasible to apply multiple imputation methods to missing continuous glucose data and make comparisons to existing methods.

P128

Efficient clinical trial designs

Multiple-treatment cluster randomised crossover trials using incomplete factorial designs

Andrew Forbes[1], Rhys Bowden[1], Jessica Kasza[1]

[1] *Monash University, Melbourne, Australia*

Cluster randomised crossover trials are a form of cluster randomised trials in which each cluster is randomised to a sequence of interventions to be delivered sequentially. The simplest case is the two-treatment-two period design: each cluster is randomised to receive either the intervention for a defined time period and then usual care, or vice versa, possibly with an interim washout period. These designs can be extended to assess multiple treatments using classical ‘row-column’ experimental designs such as Latin square designs, in which the rows of the design are clusters and the columns are periods. However, significantly more complex designs are needed in application areas such as intensive/critical care in which there are multiple aspects of the management of the physiological safety of patients. For example, one possible experiment could involve the following 5 factors: (1) daily protein intake (high/low), (2) magnesium supplementation (yes/no), (3) phosphate supplementation (yes/no), (4) Thiamine (yes/no), and (5) Vitamin C (yes/no). Each factor could be investigated in a separate cluster crossover trial, however major improvements in efficiency can be achieved by studying the 5 factors simultaneously in a 2x2x2x2x2 factorial design. A full factorial crossover design would require 2^5= 32 periods per cluster, which is not feasible. However, using experimental design ideas from agriculture, such a trial is possible using an incomplete factorial row-column design in batches of 8 clusters with only 4 observation periods per cluster. Such designs enable all main effects to be estimated efficiently using within-cluster comparisons, together with the ability to target specific first-order interactions for more efficient estimation than other interactions.
In this poster we describe these incomplete factorial designs, how they are generated, and provide guidance for mixed model analyses of these designs together with sample size estimation procedures for detecting main effects and first order interactions. We apply these principles to propose a design for the above-mentioned incomplete 2^5 factorial cluster crossover trial in intensive care.

Poster Presentations

P129

Meta-analysis

Network meta-analysis of behavioral economic incentive programs on diet, weight & physical activity

Suparee Boonmanunt[1], Oraluck Pattanaprateep[1], Boonsong Ongphiphadhanakul[1], Gareth McKay[2], John Attia[3], Ivo Vlaev[4], Ammarin Thakkinstian[1]

[1] Mahidol University, Thailand, [2] Queen's University Belfast, United Kingdom, [3] University of Newcastle, Callaghan, Australia, [4] Warwick Business School, Coventry, United Kingdom

BACKGROUND: Healthy diet, weight control and physical activity to reduce obesity can be motivated by financial incentives. Behavioral economic approaches may improve the effectiveness of those incentives. This network meta-analysis compares and ranks the effectiveness of standard and behavioral incentivization for healthy diet, weight control, and physical activity promotion.
METHODS: A systematic search of Medline and Scopus was performed from database inception to December 2020. Eligibility criteria included intervention comparisons reporting outcome goals to promote healthy diet, weight control and physical activity; intervention arms included no financial incentive (no FI), standard financial incentive (standard FI), deposit contract (deposit), and lottery-based incentive (lottery). Study characteristics, incentive program designs, and risk ratio (RR) were extracted by two reviewers. RRs were pooled using random-effects meta-analysis. Pre-specified subgroup analysis was performed. A two-stage network meta-analysis pooled and ranked intervention effects. The Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) guideline was followed.
RESULTS: There were 35 eligible RCTs. For diet-weight control, standard FI, deposit, lottery and standard FI +lottery, all increased goal achievement compared to no FI with pooled RRs and (95% confidence intervals [CI]) of 1.21 (0.94, 1.56), 1.79 (1.04, 3.05), 1.45 (0.99, 2.13) and 1.73 (0.83, 3.63), respectively. For physical activity, standard FI, deposit, and lottery, significantly increased goal achievement compared to no FI, with pooled RRs of 1.38 (1.13, 1.68), 1.63 (1.24, 2.14) and 1.43 (1.14, 1.80), respectively. In a follow-up period for physical activity, standard FI, deposit, and lottery increased goal achievement compared to no FI, with pooled RRs of 1.11 (0.94, 1.30), 1.39 (1.11, 1.73) and 1.05 (0.83, 1.32), respectively.
CONCLUSION: Deposit, followed by lottery, were best for motivating healthy diet, weight control and physical activity at program end. Post-intervention, deposit, followed by standard FI, were best for motivating physical activity. Behavioral insights can improve the effectiveness of financial incentives, although lottery-based approaches may offer only short-term benefit.

P131

Communicating statistical methods, Medical device clinical studies

Novel ways of exploring glucose patterns based on long-duration CGM data from a diabetes registry

Petra Baumann[1], Daniel Hochfellner[1], Monika Cigler[1], Laura Roubik[1], Tina Pöttler[1], Julia K. Mader[1]

[1] Medical University of Graz, Austria

Continuous glucose monitoring (CGM) provides more detailed information on glycaemic control than blood glucose (BG) measurements. Pooled analysis of long-duration CGM (>17 weeks) from a diabetes registry allows for exploration of glucose patterns in novel ways. Special attention is given to clinical usefulness and ease of interpretation of results by clinicians.
66 patients with T1D provide 24.530 days of CGM measurements (median 365.7 days per patient). We apply mixed models to account for the nested data structure and use empty models for variance decomposition. Outcomes are glucose measurements, time in range (TIR: 70-180 mg/dl) and number and mean duration of sensor-detected hypoglycaemic episodes (SDHE: <70mg/dl for >15 minutes). For glucose, we use a three-level model (measurements nested in days nested in patients) and explore the impact of daytime (night: 0-6am, day: 6-0am) and hour of the day (24 whole hours per day). For the other outcomes, we use two-level (generalized) mixed models (days nested in patient) and focus on variance decomposition. Additionally, we explore data and results graphically.
ICCs for glucose in the three-level model are 0.20 (day) and 0.15 (patient), respectively. By including hour of the day as a non-numerical fixed effect, we are able to non-linearly trace glucose over the course of the day. We plot these fixed effects over the raw data and find that the modelled curve is well-aligned with the general pattern in the data. For clinicians, interpretation of the illustration is straightforward and expedient. Fixed effect for daytime shows expectable lower glucose during the night. Fitting random slopes on both the patient and the day level shows slightly better model fit than models with random intercept only or random slope on the patient level only. This indicates that day-night differences vary both between and within patients. ICC (patient) is 0.41 for TIR, 0.04 for number of SDHE and 0.05 for mean duration of SDHE. While there is large variability between patients with regard to TIR, hypoglycaemia appears to be similar between patients, but shows larger day-to-day variability (within patients). Further analyses will aim at explaining the differences in variability on the respective levels.

P132

Simulation studies, Meta-analysis

Outlier study detection in a meta-analysis of clinical trials

Chikêola Ladekpo[1,2], Catherine Legrand[3], Céline Bugli[4], Brigitte Cheuvart[2]

[1] UCLouvain, Belgium, [2] GSK, 3 Institute of Statistics, Biostatistics and Actuarial Sciences (ISBA), UCLouvain, Belgium, 4 SMCS, UCLouvain, Belgium

For the life cycle management of a vaccine and to leverage on existing historical data, GSK seeks to improve the statistical methods for clinical monitoring including both historical and new clinical trials data. The new methodology would involve complex multivariate statistical analyses to determine whether the results of a new clinical study are in line with historical data or if they are outliers. This is of particular interest in the scope of the post-marketing surveillance of immune response to a vaccine.
In the context of a meta-analysis of vaccine clinical trials, a mixed effect model is fitted to the data to predict the antibody geometric mean titer. Based on a literature review, three promising methods were selected to detect whether a new study appears as an outlier in such meta-analysis with respect to the response variable: (1) the variance shift outlier model (VSOM), (2) the external studentized residual (ESR) and (3) the robust mixture model (RMM). These methods are compared both theoretically, via a simulation study and their application on real data.
In this presentation, we will show the results of our simulations study comparing these three approaches and the impact on the detection of the outlier study size, the amplitude of its deviancy, and the number of outlier studies in the meta-analysis. We will also present and discuss the results obtained with these three methods on a real meta-analysis database.
The ESR method is simpler, faster to implement, and more accurate given the simulated data drawn from a normal distribution. The RMM and VSOM methods involve estimation of additional variance parameters thus require more data than the ESR method for their implementation. They also include a parametric bootstrap test which makes them slower and more tedious to implement. From our preliminary results, it appears that the RMM method is not appropriate to our context as it is pushed to its limit of detecting a single outlier study as opposed to a group of outlier studies. The three methods are implemented on aggregated data, which leads to a loss of information compared to the available individual patient data.

P133

Joint models for high dimensional data

Pairwise Fitting of Piecewise Mixed Models for the Joint Modeling of Multiple Longitudinal outcomes

Moses Mwangi[1], Edmund Njeru Njagi[2], Geert Verbeke[3], Samuel Mwalili[4], Geert Molenberghs[5]

[1] Kenya Medical Research Institute, Nairobi, Kenya, [2] London School of Hygiene & Tropical Medicine, United Kingdom, [3] KU Leuven, Belgium, [4] Jomo Kenyatta University of Agriculture and Technology, Juja, Kenya, [5] University of Hasselt, Belgium

Many statistical models have been proposed in the literature for the analysis of longitudinal data. One may propose to model two or more correlated longitudinal processes simultaneously, with a goal of understanding their association over time. Joint modeling is required and gives more efficient inference than separate analyses when interested in the association structure among the outcomes or when interested in drawing joint inferences about the different outcomes. In this study, we sought to establish the associations among four nutrition outcomes while circumventing the computational challenge posed by their clustered and high dimensional nature. We analyzed data from a 2-arm, double blind, randomized cross-over trial to compare the effect of high-dose and low-dose iodine in household salt on systolic blood pressure (SBP) and diastolic blood pressure (DBP) in women of reproductive age conducted in Kenya between 22-Oct-2013 and 29-Nov-2013. Two additional outcomes namely, square root of urinary sodium concentration (sqrtUNaC) and urinary iodine concentration (sqrtUIC) were monitored. We used the pairwise joint modeling strategy to fit a correlated random effects joint model for the four outcomes. This entailed fitting 6 bivariate general linear mixed models and deriving inference for the joint model using pseudo-likelihood theory. We analyzed the outcomes separately and jointly using piecewise linear mixed model. High-dose iodine in salt significantly reduced blood pressure compared to low-dose iodine in salt. Results of the random effects showed that SBP and DBP had strong positive correlation, with effect of the random slope indicating that closely related outcomes have stronger association in their evolutions. There was a moderately strong inverse relationship between evolutions of sqrtUIC and blood pressure. These confirm the original hypothesis that high-dose iodine salt has significant lowering effect on blood pressure. In both pairwise joint modeling and univariate analysis, high-dose iodine salt reduced patients SBP and DBP overtime. The strength and direction of association varied between blood pressure and urinary excretion outcomes. Additionally, piecewise linear mixed model improved precision of parameter estimates compared to commonly used methodology.

Poster Presentations

P134 Miscellaneous

Pattern Identification in Biomedical Markers of a Mixed Type

Jan Vávra[1]

[1] *Faculty of Mathematics and Physics, Charles University, Czech Republic*

In majority of clinical studies, the patients are monitored for a longer period of time, during which many tests and examinations are gradually repeated. This creates a longitudinal dataset of diverse biomedical markers. Consider, for example, the well known PBC (primary biliary cholangitis) dataset gathered by the Mayo Clinic between 1974 and 1984. Data for this randomized placebo controlled trial of the drug D-penicillamine contain precise concentration values of albumin, bilirubin, etc. in a blood sample. Moreover, several binary indicators (presence of hepatomegaly, etc.), ordinal outcomes (seriousness of edema) or even count variables (platelet count) are also recorded. As the time progresses, each of these closely related markers evolves in relation with the overall health of the patient. Our task is to identify groups of patients that share the same evolution pattern in order to construct a classification rule for newly observed patients. First, we restrict ourselves to the patients still alive and followed after 2.5 years of the study (n=260) to imitate a real situation when data from several time points are already available, while the survival itself remains unknown and a prognosis for each patient is yet to be assessed. Generalized linear mixed-effects models (GLMM) of suitable families for diverse markers of interest are joined together through joint distribution of random effects to address and capture possible relationships among the markers. Viewed in a Bayesian setting, we elegantly overcome the problem of unobserved group allocations, latent parameters and missing data by adopting model-based clustering (MBC) and Bayesian data augmentation (BDA) principles. Carefully constructed Markov chain enhances removal of redundant groups to identify the apriori unknown number of groups. Finally, the patients are divided into the two discovered groups of contrasting patterns, which also differ in the survival potential as the data past 2.5 years suggest. Hence, a prognosis for a newly observed patient can be established by classification into one of the discovered groups.

P135 Prediction models, Personalized medicine

Performance metrics for models predicting individualized treatment effect of patients

Carolien Maas[1], David van Klaveren[1]

[1] *Erasmus MC, Rotterdam, Netherlands*

OBJECTIVE: Measuring the performance of models designed to predict individualized treatment effect is challenging, because the outcomes of two alternative treatments are inherently unobservable in one patient. The C-for-benefit was proposed to measure discriminative ability. We proposed metrics of calibration and overall performance for models predicting treatment effect.
STUDY DESIGN AND SETTING: Similar to the previously proposed C-for-benefit, we defined the observed treatment effect by the difference between outcomes in pairs of matched patients. Thus, we redefined the E-statistics, the logistic loss and the Brier score into metrics for measuring a model's ability to predict treatment effect. In a simulation study, the metric values of deliberately perturbed models were compared to those of the data generating model. To illustrate the performance metrics, different models predicting treatment effect were applied to the data of the Diabetes Prevention Program: 1) a risk modelling approach with restricted cubic splines; 2) an effect modelling approach including penalized treatment interactions; and 3) the causal forest.
RESULTS: As desired, performance metric values of perturbed models were consistently worse than those of the optimal model (Eavg-for-benefit≥0.070 versus 0.001, E50-for-benefit≥0.040 versus 0.001, E90-for-benefit≥0.115 versus 0.002, log-loss-for-benefit≥0.757 versus 0.733, Brier-for-benefit≥0.215 versus 0.212). Calibration, discriminative ability, and overall performance of the three models were similar in the case study. However, there seems to be a trade-off between calibration and discrimination since better calibrated models were worse at discriminating between patients with small and large treatment effects.
CONCLUSION: The proposed metrics are useful to assess the calibration and overall performance of models predicting individualized treatment effect.

P136 Prediction models

Performance of clinical prediction models in the presence of calibration drift

David Jenkins[1], Matthew Sperrin[1], Niels Peek[1], Camilla Sammut-Powell[1], Thomas Debray[2], Mamas Mamas[3], Glen Martin[4]

[1] *The University of Manchester, United Kingdom*, [2] *Utrecht University, Netherlands,* [3] *Keele University, Newcastle, United Kingdom,* [4] *Centre for Health Informatics, University of Manchester, United Kingdom*

BACKGROUND: Clinical prediction models (CPMs) are used in preventative medicine and to support clinical decision making. Most CPMs are developed using methods where the model remains static over time and this frequently results in worsening of predictive performance over time. Dynamic prediction models (CPMs with coefficients that are subject to Bayesian updating) and varying coefficient models (CPMs where model coefficients are a function of time), have been proposed as a solution. However, their performance has not been compared with time-invariant CPMs. We aimed to assess performance of time-invariant and dynamic CPMs under a variety of calibration drift scenarios.
METHODS: We simulated scenarios with either a continuous or binary outcome, under a variety data-generating mechanisms in which the predictor-outcome associations and outcome prevalence (binary)/mean outcome (continuous) were changing over time. Time-invariant regression models, dynamic (Bayesian updating) prediction models, and varying coefficient models were fitted to the data and their predictive performance was assessed. We also applied all of the methods in a real-world cardiac surgery dataset to predict in-hospital mortality and compared the performance of the resulting models.
RESULTS: In the simulation, when no calibration drift was present, all models had similar predictive performance. When calibration drift was present, the Bayesian and varying coefficient models outperformed the time-invariant model. In the cardiac surgery data, the dynamic Bayesian model was calibrated-in-the-large, -0.018 (95% confidence interval: -0.039 to 0.002), but the time-invariant and varying coefficient models were miscalibrated-in-the-large, 0.158 (95% confidence interval: 0.137 to 0.178) and 0.158 (95% confidence interval: 0.137 - 0.178), respectively.
CONCLUSION: Dynamic models based on Bayesian updating methods improve CPM performance in the presence of calibration drift, and also perform at least as good as time-invariant CPMs when there are little to no temporal changes. Although outperforming the time-invariant model in the simulation, the varying coefficient models performance was similar to the time invariant model, and was outperformed by the Bayesian model, in the cardiac surgery data. We recommend the use of dynamic prediction models over time-invariant and varying coefficient models.

P137 Prediction models, Meta-analysis

Performance of methods for meta-analysis of incremental predictive value: a comparison study

Junfeng Wang[1], Artuur Leeuwenberg[1], Ewoud Schuit[1]

[1] *Utrecht University, Netherlands*

BACKGROUND: Novel markers can improve the prediction of diagnostic and prognostic outcomes, and the incremental value of a new marker should be assessed over a widely used prediction model. However, statistical methods in Systematic reviews and meta-analyses of incremental value of a new predictor are still underdeveloped and different strategies need to be investigated on their accuracy in estimating the true incremental value.
METHODS: We compare three different strategies in performing meta-analysis of the incremental predictive value of a marker: (1) Univariate meta-analysis of ΔAUC (defined as AUCnew - AUCref); (2) Bivariate meta-analysis of AUCref and AUCnew; and (3) Bivariate meta-analysis of AUCref and ΔAUC. The performances of these methods are evaluated via simulations with different scenarios reflecting realistic situations. We assume there are 7 predictors which follow multivariate normal distribution with mean of 0, standard deviation of 1 and correlation of ρ [ρ= 0, 0.2, or 0.4]. The true model (which generates the binary outcome) includes all 7 predictors, and the reference model only includes the first 6 predictors. We evaluate the 7th predictor's incremental predictive value, with its β changes among [0, 0.4, 0.7, or 0.9]. The true incremental value is estimated in a population of 1,000,000. The sample size in each primary study follows normal distributions [N(200, 20), N(500, 50), or N(1000, 100)], and each meta-analysis includes [5, 10, or 20] primary studies. In total, 216 scenarios are used to compare the three methods.
RESULTS: In general, the true incremental value increases when β increases and decreases when ρ increases. The performances of all three strategies increase when the number of primary studies or the sample size in each primary study increases. When the new predictor has no incremental value, method (2) always overestimates the incremental value and has highest RMSE. However, method (2) performs better when the new predictor does improve the prediction, and the advantage is more significant when the number of studies or the sample size in each primary study is smaller. The performance of method (1) and (3) are similar in all scenarios.
CONCLUSIONS: Based on the results, we would recommend method (1) or (3) when the incremental value is small, and method (2) when a clear incremental value is observed in primary studies.

Poster Presentations

P138

Simulation studies

Performance of the P30 measure for assessing the accuracy of estimating glomerular filtration rate

Katie Scandrett[1], Yemisi Takwoingi[1], Stephanie Hubbard[2], Alice Sitch[1]

[1] *University of Birmingham, United Kingdom*
[2] *University of Leicester, United Kingdom*

CONTEXT: Difficulties in measured glomerular filtration rate (mGFR) for diagnosing and monitoring chronic kidney disease (CKD) has led to the development of many equations to obtain estimated GFR (eGFR). The accuracy of eGFR equations is commonly measured using P30 (percentage of eGFR values within 30% of mGFR for a group of individuals) and yet this metric was defined with no clinical or statistical rational. Other measures looking at smaller differences, such as 5%, 10% and 15% (P5, P10 and P15), have been recommended but are less commonly used.
OBJECTIVES: To investigate whether P30 is the optimal measure to use when evaluating the accuracy of eGFR equations. P5, P10 and P15 were considered as potential alternatives. We calculate these measures in a case study (using the eGFR-C study data) and then simulate data sets with different levels of bias to further investigate the performance of P30 and its variations.
Methods: After calculating the measures in the case study, true GFR values were generated using a uniform distribution for 10,000 subjects. eGFR and mGFR were generated from these true GFR values using two different models: one where bias was modelled as absolute (eGFR-mGFR = constant) and one where bias was modelled in proportion of true GFR. Constants were incorporated into the eGFR and mGFR values to represent within-subject biological variability and analytical error.
RESULTS: P30 ranked the accuracy of the equations differently to the P5, P10 and P15 when applied to the case study. The simulation showed that the P30 is insensitive to low levels of bias and is not robust to the range of GFR values of the subjects within a specific study population. P5, P10 and P15 are less affected by these issues. Even where there was no bias present, P5, P10 and P15 were low.
CONCLUSION: Use of P30 alone could lead to the wrong conclusions being drawn and should be interpreted with caution. P5, P10 and P15 should be also considered when evaluating the accuracy of eGFR equations. Further research is due to identify alternative methods of measuring the accuracy of eGFR equations.

P139

Simulation studies, Personalized medicine

Phenotypic deconvolution in heterogeneous cancer cell populations using drug screening data

Even Moa Myklebust[1], Alvaro Köhn-Luque[1], Dagim Shiferaw Tadele[2], Mariaserena Giliberto[3], Jasmine Noory[4], Elise Harivel[1], Polina Arsenteva[1], Shannon M Mumenthaler[5], Fredrik Schjesvold[6], Kjetil Taskén[3], Jorrit Martijn Enserink[2], Kevin Leder[7], Arnoldo Frigessi[8], Jasmine Foo[4]

[1] *Oslo Centre for Biostatistics and Epidemiology, University of Oslo, Norway,* [2] *Department of Molecular Cell Biology, Institute for Cancer Research, Oslo University Hospital, Norway,* [3] *Department of Cancer Immunology, Institute for Cancer Research, Oslo University Hospital, Norway,* [4] *Institute for Mathematics and its Applications, School of Mathematics, University of Minnesota, Minneapolis MN, United States,* [5] *University of Southern California, Lawrence J. Ellison Institute for Transformative Medicine, Los Angeles CA, United States,* [6] *KG Jebsen Center for B-Cell Malignancies, Institute for Clinical Medicine, University of Oslo, Norway,* [7] *College of Science and Engineering, University of Minnesota, Minneapolis MN, United States,* [8] *University of Oslo, Norway*

Tumor heterogeneity is an important driver of tumor recurrence, as treatments that initially elicit clinical responses can select for drug-tolerant tumor subpopulations, leading to the outgrowth of resistant clones and cancer treatment failure. Profiling the drug-response heterogeneity of tumor samples using traditional genomic deconvolution methods has yielded limited results, due in part to the imperfect mapping between genomic variation and functional characteristics. Here, by introducing an underlying population dynamic model of tumor subclonal response to therapy, we enable the phenotypic deconvolution of bulk drug-response data into component subpopulations, and the estimation of their differential drug sensitivities and population frequencies. We used this method, called DECIPHER, to perform deconvolution on tumor drug screening data generated both experimentally and in silico. This study demonstrates how mechanistic population modeling can be leveraged to develop statistical frameworks for profiling phenotypic heterogeneity from bulk tumor samples and to perform individualized patient treatment predictions.

P140

Prediction models

Predicting multidrug resistance in neutropenic cancer patients with bloodstream infection

Natalia Pallares[1], Pau Satorra[1], Adaia Albasanz-Puig[2], Júlia Laporte-Amargós[2], Carlota Gudiol[2], Cristian Tebé[1]

[1] *IDIBELL - Bellvitge Biomedical Research Institute, Barcelona, Spain,* [2] *Bellvitge University Hospital, Barcelona, Spain*

OBJECTIVE: To develop a clinical prediction model to estimate multidrug resistance risk in onco-hematological patients with neutropenia and bloodstream infection (BSI) due to multidrug-resistant Pseudomonas aeruginosa (MDRPA).
METHODS: A multicenter, retrospective cohort study including neutropenic patients with BSI due to MDRPA was conducted in 34 centers (12 countries) between 2006 and 2018. The study sample was split into a derivation (80%) and a validation cohort (20%).
A mixed logistic regression model with random intercept (center) was used to develop a predictive model. Ten multiple imputations with chained equations (MICE) were used to deal with missing data. Each of the 10 datasets created was sampled by bootstrapping with replacement 100 times, totaling 1000 samples. A model was fitted in each sample using backward elimination (AIC criteria). Predictors retained in more than 70% of the 1000 estimated models were considered for inclusion in the final model. Rubin's rule was used to summarize a model with the selected predictors. Discrimination was assessed by estimating the area under de ROC curve (AUC), calibration by comparing observed versus expected MDR by tenths of predicted risk. All validation analyses performed in the derived sample was also repeated in a preserved sample for validation.
RESULTS: Among 1217 episodes of BSI (38.3% women, 57.8-yo, 75.3% hematologic disease), 25% were caused by an MDR strain (95%CI 19.7 to 30.9). After bootstrapping process elected variables were prior therapy with piperacillin/tazobactam, prior antipseudomonal carbapenem use, fluoroquinolone prophylaxis, hematological underlying disease, presence of a urinary catheter, and age. The predictive model obtained in the derivation cohort had excellent discrimination with an AUC of 0.82 (95% CI 0.79-0.85). The observed probability corresponded well to the predicted probability. The performance in the derivation and validation cohort was similar. A Shiny app (https://ubidi.shinyapps.io/ironic/) was developed to calculate the risk of MDR.
CONCLUSIONS: This prediction model will improve the identification of patients at high risk for multidrug resistance and will avoid the use of broad-spectrum antibiotics in patients with less risk of resistance.

P141

Efficient clinical trial designs, Prediction models

Prediction of clinical trial cycle times with Bayesian Model Averaging

Kleio Kipourou[1], Nathan Debellis Amaral[1], Aris Perperoglou[1], Nick Carroll[1], Hrishikesh Karvir[1]

[1] *AstraZeneca, United Kingdom*

There is growing evidence [1] that the total duration of clinical trials has been steadily increasing since 2006, with the recent COVID-19 outbreak magnifying the problem. Clinical trial cycle times are defined in pharmaceutical industry as time elapsed between study start-up and first-patient-in or patient recruitment completion. Among other factors, slowness in start-up phase and patient recruitment are considered some of the main reasons for delays in the overall drug development process [2]. The challenge in reviewing year on year cycle time performance is the relatively small number of trials completing a cycle in any given year and detecting true changes in times with the high variance related to small numbers and other factors.
In this work we analyse historical data on cycle times of 100 clinical trials sponsored by AstraZeneca from 2006 to today. We present trends in cycle times by therapeutic area. Data show that cycle times are increasing following similar patterns described in the literature.
Additionally, we present how our predictive models of recruitment can be used to predict cycle times on ongoing studies providing a solution that addresses the small number of trials completing in any given year. We implement a family of non-homogeneous Poisson models [3] and predict study recruitment with Bayesian Model Averaging. The models incorporate information from historic data and observed trends to predict future recruitment under a series of scenarios, where recruitment rate may remain constant or change with time. These models have been implemented in the last 2 years within AstraZeneca and have been used daily to monitor study performance in an increasing portfolio of studies. We present predictive performance on historical cycle times and current estimates of future cycle times in a portfolio of current active studies.

REFERENCES:
[1] Martin L, Hutchens M, Hawkins C. «Trial watch: clinical trial cycle times continue to increase despite industry efforts.» Nature Reviews Drug Discovery (2017): 157-158.
[2] Lamberti MJ, Brothers C, Manak D, Getz K. Benchmarking the study initiation process. Therapeutic Innovation & Regulatory Science (2013): 101-109.
[3] Urbas S, Sherlock C, Metcalfe P. Interim recruitment prediction for multi-center clinical trials. Biostatistics. (2022): 485-506.

Poster Presentations

P142

Prediction models, Personalized medicine

Predictive modeling approaches to personalized medicine: a comparison of regression-based methods

David van Klaveren[1], David M Kent[2], Ewout W Steyerberg[3]

[1] Erasmus MC, Rotterdam, Netherlands, [2] Tufts Medical Center, Boston MA, United States, [3] Leiden University Medical Center, Netherlands

The benefits and harms of medical treatments vary substantially between individual patients. Predictive modeling approaches to personalized medicine are designed to predict the benefit of one treatment over another for individual patients. We aimed to compare different regression-based modeling approaches, through simulations and a case-study.
We simulated trial samples (n = 3,600; 80% power for a treatment odds ratio of 0.8) from a superpopulation (N = 1,000,000) with 12 binary risk predictors, both without and with six true treatment interactions. We assessed predictions of treatment benefit for four regression models: a «risk model» (with a constant effect of treatment assignment) and three «effect models» (including interactions of risk predictors with treatment assignment). The risk modeling approach was well-calibrated for treatment benefit, whereas effect models were consistently overfit, even with doubled sample sizes. Penalized regression reduced miscalibration of the effect models considerably. In terms of the benefit prediction error, the risk modeling approach was superior in the absence of true treatment effect interactions, whereas penalized regression was optimal in the presence of true treatment interactions.
The recently proposed Syntax Score II (SSII)-2020 was developed to predict the difference in 10-year mortality when treating complex coronary artery disease patients with heart bypass surgery rather than coronary stenting. Cox regression was first used in the SYNTAX trial data (n=1,800) to develop a prognostic index (PI) for mortality over a 10-year horizon consisting of 7 clinical predictors of mortality. Second, a Cox model was fitted which included the treatment, the PI and pre-specified treatment interactions with type of disease and with anatomical disease complexity. In contrast to its more flexible predecessor SSII-2013 which included 8 treatment interactions, SSII-2020 was well calibrated for treatment benefit at 10 years post-procedure, both at cross-validation in the same data and at external validation in new data.
The simulations and the case study both showed that robust modeling approaches – only including plausible treatment interactions – may lead to better predictions of treatment effect. Future research could focus on robust approaches for data-driven selection of treatment interactions.

P143

Statistical education

Publishing your book with CRC

Lara Spieker[1]

[1] Chapman & Hall/CRC

In this very practical poster, an editor from Chapman and Hall/CRC discusses how to publish a statistics book with us. The poster will go over the publishing process and provide best practices for shaping your ideas and submitting a book proposal; the editor will discuss the bestsellers and popular series, as well as emerging topics and trends.

P144

Simulation studies, Missing data

Quantifying the problem of inconsistent missing data handling: A simulation study

Antonia Tsvetanova[1], Matthew Sperrin[1], Niels Peek[1], Iain Buchan[2], Stephanie Hyland[3], Glen Martin[4]

[1] The University of Manchester, United Kingdom, [2] The University of Liverpool, United Kingdom, [3] Microsoft Research, Cambridge, United Kingdom, [4] Centre for Health Informatics, University of Manchester, United Kingdom

No clear guidance exists on handling missing data at each stage of clinical prediction models (CPMs) development, validation and implementation. Although attempts have been made to identify which missing data approaches are consistent by us and others, this has not been examined empirically. This simulation study aims to determine whether the performance of a CPM estimated in validation sufficiently represents the performance one would see in implementation for a logistic regression-based CPM.
Simulations were designed to mimic each of the stages of a CPM's pipeline – development, validation and implementation. We divide the data generation into two main steps – the first is mimicking a development dataset and the second – external validation and implementation data. The development data is fully observed, with X1, X2, U and outcome Y, whilst the validation and implementation cohorts contain missing data in X1, denoted by R1, ranging from 5% to 50%. We construct five scenarios in which the data is missing, covering missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). We fit our initial model to the fully observed development data, which serves as a reference for perfect calibration and optimal discrimination of the 'true' model. We obtain parameters from the fully observed model. We apply three models to the validation and implementation data, which are either miscalibrated or with poor discrimination and of which, we will be assessing the performance. Missing data is imputed using four different missing data handling approaches – multiple imputation, complete case analysis, mean imputation, zero imputation (for binary variables).
We predict the risk based on the models (i) – (iv), then we calculate the performance statistics. Our key target is the predictive performance of an individual's predicted risk, and therefore we are interested in estimating the standard errors, the coverage of the CI, and the bias of the following estimands, covering both calibration and discrimination: Mean squared error, calibration in the large, calibration slope, O-E ratio, C-statistic.
We will compare the 'ground truth' defined as the performance at the implementation stage to the estimations at validation and we will determine which of these methods' combinations lead to under- or over- estimates of the CPMs performance.

P145

Personalized medicine

Quantitative decision-making in the context of early-phase biomarker-adaptive designs

Amelia Thompson[1], Nigel Stallard[1], Deepak Parashar[1], Paul Frewer[2]

[1] University of Warwick, Coventry, United Kingdom, [2] AstraZeneca, United Kingdom

In recent years a number of decision-making frameworks have been proposed as an alternative to the Null Hypothesis Significance Testing (NHST) approach. The three-outcome framework from Frewer et al (2016), based on Lalonde et al (2007) utilises two pre-set values, namely the Target Value and Lower Reference Value, by which to judge the estimated outcome and categorize it as either a go, consider or stop decision.
The work presented here builds upon this framework, introducing interim analysis, adaptive elements and biomarker stratification. We specifically focus on the early phase oncology setting. The designs are evaluated through simulation and approximation, achieving the desired operating characteristics. Such operating characteristics include false go and false stop rates, analogous to type I and type II errors of NHST as well as the probability of a consider decision and expected sample size.
The novel designs developed in our work maintain a high level of control over the probability of an incorrect decision while also allowing the possibility of a smaller sample size. It also encourages actions to be clearly predefined for each outcome. This allows for quick unbiased decision making either at the interim or the end of a trial.

Poster Presentations

P146

Joint models for high dimensional data, Meta-analysis

Quirks with joint hierarchical models: Examples involving global relationships between sodium and potassium intake, GDP and healthy life expectancy

John Ferguson[1], Alberto Alvarez[1], Catriona Reddin[1], Martin O'Donnell[1], Robert Murphy[1]

[1] HRB Clinical Research Facility, NUI Galway, Ireland

We describe a joint hierarchical Bayesian model for regional and country-level per-capita potassium and sodium intake. The model is fit to data collected from 137 separate studies pertaining to 52 countries. Extending this joint Bayesian model to include a second level regression for country-level healthy life expectancy strongly affected some country and region level estimates for sodium and potassium intake. Accounting for spatial correlation in healthy life expectancy lessened this effect somewhat. These results raise the question of whether including such second level regressions (as a form of indirect evidence) is advised in such analyses when meta-analysis of country-level means or prevalences is of primary interest

P148

Miscellaneous

Recommendations for evaluating and regulating in vitro diagnostic tests

Katie Scandrett[1], Deborah Ashby[2], Jon Deeks[1], Sheila Bird[3], Stephen Evans[4], Rafael Perera[5], Yemisi Takwoingi[1]

[1] University of Birmingham, United Kingdom, [2] Imperial College London, United Kingdom, [3] University of Cambridge, United Kingdom, [4] London School of Hygiene and Tropical Medicine, United Kingdom, [5] University of Oxford, United Kingdom

During the COVID-19 pandemic, testing has been a key pillar of the UK government's strategy to protect the NHS and save lives. The RSS were concerned that many new diagnostic tests for SARS-CoV-2 antigen or antibodies were coming to market for use both in clinical practice and for surveillance without adequate provision for statistical evaluation of their analytical and clinical performance.
The RSS Diagnostic Tests Working Group reviewed the statistical evidence needed to assure the performance of new diagnostic tests, for patients, decision-makers and regulators, with particular reference to in-vitro diagnostics (IVDs) for infectious diseases. Recommendations were developed in three key areas: study-design, regulation and transparency.
Key issues regarding study-design include: carefully defining the intended use of an IVD; incorporating knowledge of disease prevalence into studies; inclusion of both field and analytical performance study-designs; presenting all estimates with a measure of uncertainty; clearly stating the assumptions of mathematical models and improving planning for future pandemics so that future studies can commence effectively. In terms of regulation, the impact of tests beyond physical safety must be considered, including the consequences of any false positive or negative results and any unintended consequences arising from use of the test. Reference standards must be agreed and regulators should revise the national licensing process and create Target Product Profiles for IVDs. Transparency of studies reviewing IVDs is vital and should be carried out with the same rigour as randomised-control trials.

P149

Missing data, Statistical software

Reconstructing KIR haplotypes based on ambiguous genotype data and outcome models

Lars van der Burg[1], Stefan Böhringer[1], Henning Baldauf[2], Jürgen Sauter[2], Johannes Schetelig[2,3], Hein Putter[1], Liesbeth C. de Wreede[1,2]

[1] Biomedical Data Sciences, LUMC, Leiden, Netherlands , [2] DKMS, Dresden/Tübingen, Germany, [3] Department of Internal Medicine I, University Hospital Carl Gustav, Dresden, Germany

With the advances of DNA sequencing technologies, an enormous amount of unphased genotype data is obtained. Arguably, the analysis of such data is best exploited through phased haplotypes. Haplotypes frequencies can be estimated via the Expectation-Maximation (EM)-algorithm, which treats the unobserved haplotype phase as. To keep the EM-algorithm computationally feasible, the number of HTs needs to be kept low.
For certain genes this limitation can be problematic. The KIR (killer-cell immunoglobulin-like receptors) gene region is genetically complex: exhibiting copy number variations, huge allelic diversity and genotypes are often measured with ambiguities. These KIR genes are hypothesized to play a beneficial role after allogeneic hematopoietic stem cell transplantation. However, for clear conclusions haplotypes need to be reconstructed.
Previously, we presented an EM-algorithm that dealt with the limitations. However, this algorithm only used the observed genotype information. For imputation of missing predictors it is in general necessary to include the outcomes. Horton & Laird (1998) described how the outcome can be included in a general EM-algorithm for the exponential family distribution via the GLM framework, which we implemented as an extension in our EM-algorithm. By fitting regression models in each iteration, haplotype effects are estimated and subsequently used to update their estimated frequencies. Modelling choices are required to limit the extensive number of haplotypes, three different strategies are developed.
The first reduces all haplotypes into allelic "main" effects for each gene. This greatly reduces the dimensionality, but cannot incorporate haplotype effects. The second method continues with the main effects, and includes haplotypes via a forward-backward selection procedure. However, the choice for specific haplotypes can differ per iteration and influence the rest of the algorithm. As third strategy all possible main effects, lower- and higher-level haplotypes are included using penalized regression. This allows for separate effects, but due to the penalty these will be biased. Although our method is general, there is no 'one-size-fits-all' approach possible, with strategy choice depending on the assumed biological effect pathway.

P150

Prediction models

Regularized parametric survival modeling to improve risk prediction models

Jeroen Hoogland[1], Thomas Debray[1], Michael Crowther[2,3], Richard Riley[4], Joanna intHout[5], Johannes Reitsma[1], Aeilko Zwinderman[6]

[1] Julius Center, University Medical Center Utrecht, Utrecht University, Netherlands, [2] Red Door Analytics, Stockholm, Sweden, [3] Karolinska Institute, Stockholm, Sweden, [4] Keele University, Newcastle, United Kingdom, [5] Radboud Institute for Health Sciences (RIHS), Radboud University Medical Center, Nijmegen, Netherlands, [6] Clinical Epidemiology and Biostatistics, Amsterdam University Medical Center, Netherlands

We propose to combine the benefits of flexible parametric survival modeling and regularization to improve risk prediction modeling in the context of time to event data. Thereto, we introduce ridge, lasso, elastic net, and group lasso penalties for both log hazard and log cumulative hazard models. The log (cumulative) hazard in these models is represented by a flexible function of time that may depend on the covariates (\textit{i.e.}, covariate effects may be time-varying). We show that the optimization problem for the proposed models can be formulated as a convex optimization problem and provide a user-friendly R implementation for model fitting and penalty parameter selection based on cross-validation. Simulation study results show the advantage of regularization in terms of increased out-of-sample prediction accuracy and improved calibration and discrimination of predicted survival probabilities, especially when sample size was relatively small with respect to model complexity. An applied example illustrates the proposed methods. In summary, our work provides both a foundation for and an easily accessible implementation of regularized parametric survival modeling and suggest that it improves out-of-sample prediction performance.

Poster Presentations

P151

Miscellaneous

Repeatability of Dynamic Contrast Enhanced Magnetic Resonance Imaging (DCE-MRI): A Systematic Review

Nivetha Sridharan[1], James PB O'Connor[1], Nuria Porta[1]

[1] *The Institute of Cancer Research, London, United Kingdom*

BACKGROUND: Dynamic contrast enhanced magnetic resonance imaging (DCE-MRI) is used widely to assess response to cancer therapy, both in clinical practice and in early phase trials. The most common derived measurement is Ktrans - a quantitative imaging biomarker (QIB) that reflects blood flow and vessel permeability. Measurement repeatability can help identify the magnitude of true change at the individual and population level and is usually estimated in test-retest studies. However, reporting of Ktrans repeatability is seldom done, and when performed it occurs in small studies and with metrics that are often not generalisable . Here, we conduct a systematic review to determine the extent of literature reporting of Ktrans repeatability in cancer.
METHODS: A systematic search of databases (Pubmed, Web of Science and Cochrane) was performed to identify DCE-MRI test-retest studies in cancer patients reporting Ktrans repeatability. Identified studies were screened independently by two authors (NS, NP) and discussed by all authors. Study information as well as the commonly used repeatability metrics were extracted when reported: coefficient of variation (CV), repeatability coefficient (RC) and associated limits of agreement (LOA), and intra-class correlation coefficient (ICC). This study was performed by following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.
RESULTS: After screening 301 articles, 50 duplicate studies were removed and 19 eligible articles reporting the repeatability of Ktrans were retained. Among the 19 studies, the sample size ranged from 9 to 44, and 6 studies (32%) reported only one repeatability metric. CV was reported in 16 (80%), and its values ranged from 7.7%-29%; RC were reported in 9(45%), and its values ranged 0.03-72; 6 (30%) studies reported ICC, with values ranging from 0.686-0.982.
CONCLUSIONS: There is great variation in the reporting of repeatability for DCE-MRI Ktrans. Metrics such as ICC do not lend themselves to meta-analysis as it depends on the study-specific variance. Further work will evaluate individual patient data meta-analysis to robustly estimate its RC so true change in Ktrans can be defined.

P152

Personalized medicine, Meta-analysis

Reporting of sequential multiple assignment randomized trial design studies: a systematic review

Theophile Bigirumurame[1], Germaine Uwimpuhwe[2], James Wason[1]

[1] *Newcastle University, Newcastle upon Tyne, United Kingdom,* [2] *Durham University, United Kingdom*

A sequential multiple assignment randomisation trial (SMART) is a multistage trial design that can be used to develop and refine effective AIs, also known as dynamic treatment regimens (DTR), adaptive treatment strategies (ATS), multistage treatment strategies (MTS) or treatment policies. One of the appealing features of the SMART design is the possibility to re-randomise all or a group of patients, mainly non-responders to the initial treatment, based on the response, patient's characteristics, or behaviours observed during the previous treatment.
We systematically searched four databases (PubMed, Ovid, Web of Science, and Scopus) for all trial reports, protocols, reviews, and methodological papers which mentioned the SMART, in order to assess the quality of reporting of the information required to design Sequential Multistage Assignment Randomised Trial (SMART) studies. We did not put time restrictions during our search.
Of the 157 selected records, 12 (7.64%) were trial reports, 24 (15.29%) were study protocols. All these trials were powered using stage-specific aims. Only four (33.33%) of these trials reported parameters required for sample size calculations. A small number of the trials (16.67 %) were interested in determining the best embedded adaptive interventions. Most of the trials did not report information about multiple testing adjustments.
Furthermore, most of the records reported designs that were mainly focused on stage-specific aims.
Despite the increase in the use of SMART design, there are still some appealing features of this design that are not widely used or reported well. Furthermore, studies using this design tend to not adequately report information about all the design parameters.

P153

Simulation studies, Medical device clinical studies

Resampling methods for the agreement ICC in the case of non-normality with small sample size

Abderrahmane Bourredjem[1], Hervé Cardot[2], Hervé Devilliers[3]

[1] *INSERM CIC1432, Dijon University Hospital, France,* [2] *IMB - Institut de Mathématiques de Bourgogne, France,* [3] *Internal Medecine and Systemic Diseases Unit, Dijon University Hospital, France*

The agreement intra-class correlation coefficient (ICCa), issued from a two-way crossed random effects model (subject*rater), is a reliability index recommended by the regulatory agencies for quantitative measures validation and is the only parameter allowing the reliability generalization to the whole raters population. Nevertheless, its estimators are biased and many solutions have been tried facing to its confidence interval (CI) problem. The latest works indicate that no method works well with a hard-to-detect violations of normality and when the number of subjects OR raters is limited, which is rather the case in practice, especially for the raters number. This study proposes to evaluate the contribution of resampling methods to improve this weakness. Methods: Resampling methods were adapted from the generalizability theory. We explored two Jackknife strategies: a delete 1-patient and a delete 2-patients as well as two bootstrap ones: patients-cluster (Boot-p) and random-effect predictors (Boot-p,i,r), with variance components bias-adjustment and using percentile, bias-corrected accelerated (BCa) and studentized CI formulas. Logarithmic and an approximate variance-stabilizing transformation (VST) were also applied. These methods were compared by simulations to the most relevant existing methods: Fleiss & Shrout, the modified large sample and to a recent F and beta-distribution approximations using the REML estimators. Scenarios include normal and several non-normal distributions with 3 to 6 raters and 20 to 40 patients. An illustration will be provided based on orthopedic imaging real data. Results: We proof that the VST is not identifiable for the ICCa and propose an approximated transformation making its dispersion nearly constant (DNCT). Simulations show that the delete 2-patients and Boot-p resampling strategies were not appropriate. Jackknife delete 1-patient improves the bias. In case of non-normality, Boot-p,i,r improves the coverage rate of CI depending on the underlying data distribution. The DNCT improves the coverage rate of the delete 1-patient CI but is still inferior to the expected level. Conclusions: Jackknife delete 1-patient and Boot-p,i,r resampling strategies would be an interesting alternatives for the ICCa inference in the case of non-normality or few raters. However, the optimal CI formula seems depend on the data distribution.

P154

Prediction models, Personalized medicine

Restricted mean survival time regression model with time-dependent covariates

Chengfeng Zhang[1], Baoyi Huang[1], Hongji Wu[1], Yawen Hou[2], Zheng Chen[1]

[1] *Southern Medical University, Guangzhou, China,* [2] *Jinan University, Tripoli, Lebanon*

In clinical or epidemiological follow-up studies, methods based on time scale indicators such as the restricted mean survival time (RMST) have been developed to some extent. Compared with traditional hazard rate indicator system methods, the RMST is easier to interpret and does not require the proportional hazard assumption. To date, regression models based on the RMST are indirect or direct models of the RMST with baseline covariates. However, time-dependent covariates are becoming increasingly common in follow-up studies. To address survival data with time-dependent covariates and time-effect covariates, we first proposed time-dependent RMST (T-RMST) regression based on the IPCW method, which is handled exogenous time-dependent covariates. Further to deal with endogenous time-dependent covariates (longitudinal covariates), we proposed joint models of RMST and longitudinal covariates (JM-RMST) on the basis of the T-RMST models, which can dynamically predict individual survival time. Through Monte-Carlo simulation, We verified the estimation performance of the regression parameters of the T-RMST model, and the prediction performance is better than time-dependent Cox and fixed covariate RMST regression. Besides, JM-RMST regression coefficient estimates are also stable, and the prediction performance is better than the fixed covariate RMST regression. Finally, time-dependent RMST regression and JM-RMST regression are illustrated by two examples.

Poster Presentations

P157

Communicating statistical methods

Sample size estimation in RCT with annualized relapse rates can be improved: a review

Bruno Pereira[1], Louis Poncet-Megemont[1], Fabien Rollot[2], Maria Pia Sormani[3], Pierre Clavelou[1], Xavier Moisset[1]

[1] University Hospital Clermont-Ferrand, France, [2] Lyon University, France , [3] University of Genoa and Ospedale Policlinico San Martino, Italy

Luckily for patients, relapses are generally rare. The simplest way to model rare events is to use the Poisson distribution. The quasi-likelihood Poisson distribution and negative binomial (NB) regression are Poisson derived distributions that allow overdispersion of the data. Overdispersion results from heterogeneity of the relapse rate between patients and contagion, both mechanisms leading to a statistical link between relapses. Thus, these methods are both suitable to study ARRs and are generally used to analyze the results in multiple sclerosis (MS) studies, although NB regression is preferable. These methods should also be used for SSE, which is infrequently the case. To determine whether the sample-size estimation (SSE) and the analysis of annualized relapse rates (ARR) in randomized controlled trials (RCT) were aligned and compare the SSE between normal and NB distributions.
Systematic review of phase 3 and 4 RCTs for which the primary endpoint was ARR in relapsing remitting MS published since 2008 in pre-selected major medical journals. We checked whether the SSE and ARR analyses were congruent. We also performed standardized (fixed α/β, number of arms and overdispersion) SSEs using data collected from the studies.
Twenty articles (22 studies) were selected. NB distribution (or quasi-Poisson) was used for SSE in only 7/22 studies, whereas 21/22 used it for ARR analyses. SSE relying on NB regression necessitated a smaller sample size in 21/22 of our calculations. SSE was rarely performed using the most appropriate model. For MS RCTs studying relapses, SSE based on NB distribution is recommended. This method is available and it is methodologically necessary to use the same method for both planning of the study and analysis of the results. The use of such an appropriate method optimizes the number of patients to be included, which is necessary from both an ethical and economic point of view. Indeed, SSE based on NB distribution tends to decrease the number of participants per treatment arm and clinicians must be aware that future studies may include fewer subjects than in the past, while retaining appropriately powered results. MS studies should avoid using a normal distribution for SSE when the relapse rate is the primary outcome to ensure alignment between the sample size calculation and analytical method and the rigour of the study design.

P158

Machine learning methods for health

Screening models for chronic periodontitis

Htun Teza[1], Anuchate Pattanateepapon[1], Attawood Lertpimonchai[2], Ammarin Thakkinstian[1]

[1] Mahidol University, Bangkok, Thailand, [2] Chulalongkorn University, Bangkok, Thailand

BACKGROUND: Chronic periodontitis affects 11.2% of the population globally, resulting in loss of teeth and reduced quality of life. Although periodontal probing is the gold standard, it is time- and resource-consuming, so screening models to identify high risk of chronic periodontitis can help provide more targeted resourcing. Performance of traditional statistical performance depends on optimal feature selection and a priori hypotheses. While machine learning approaches have recently gained favour given their potential discriminatory power and no a priori hypotheses requirements, most classification models are limited to independent and identically distributed random variables assumption. With our study, we applied statistical and machine-learning approaches tailored to learn from repeatedly measured data and compared the performance of disease risk prediction amongst them.
METHODS: 1817 subjects followed up with 5-years-interval were used to develop mixed-effects logistic regression (MELR), recurrent neural networks (RNN), and mixed-effects support vector machine (ME-SVM) models. Outcomes are labelled using Centre for Disease Control – American Academy of Periodontology defined criteria, and models are evaluated using uniform decision-threshold metrics.
RESULTS: The MELR model (90.5% accuracy, 86.9% F-score) performed better than the machine learning approaches (RNN: 70.0% and 57.3%; ME-SVM: 72.7% and 56.4%) to identify those with severe chronic periodontitis.
CONCLUSION: MELR models could be applied to large-scale screening of public health missions or electronic health records to identify those at highest risk of chronic periodontitis following independent external model validation.

P160

Meta-analysis

Sexual dysfunction between laparoscopic and open inguinal hernia repair: A meta-analysis

Napaphat Poprom[1], Pichet Wattanapreechanoni[1], Sitanun Saengsri[1], Chairat Supsamutchai[1], Chumpon Wilasrusmee[1]

[1] Department of Surgery, Faculty of Medicine, Ramathibodi Hospital, Mahidol University, Bangkok, Thailand

BACKGROUND:Sexual dysfunction after inguinal hernia complication is considered rare, however, its consequences impact one quality of life inevitably. Laparoscopic and open inguinal hernia repair may comparable in terms of recurrent rate, overall complication and chronic pain. The sexual dysfunction complication is still questionable between these approaches. We aimed to compare sexual dysfunction and related complication between laparoscopic and open inguinal hernia repair.
METHODS: Systematic review and meta-analysis of randomized controlled trials (RCTs) compared laparoscopic and open inguinal hernia repair. Risk ratio (RR) and 95% confidence intervals (CI) were used as pooled effect size measures.
RESULT: Thirty RCTs (12,022 patients) were included. Overall, 6,014 (50.02%) underwent laparoscopic hernia repair and 6,008 (49.98%) underwent open hernia repair. Laparoscopic approach provided non-significance benefit on pain during sexual activity (RR 0.57; CI 0.18-1.76), vas deferens injury (RR 0.46; CI 0.13-1.63), testicular pain (RR 1.37; CI 0.81-2.31), orchitis (RR 0.84; CI 0.61-1.17), scrotal hematoma (RR 0.99; CI 0.62-1.60) and testicular atrophy (RR 0.46; CI 0.17-1.20). Meanwhile, open inguinal hernia approach reduced cord seroma complication (RR 1.79; CI 1.13-2.86)
CONCLUSION: There is no advantage of laparoscopic inguinal hernia repair over open approach in term of sexual dysfunction. On the contrary, there is increasing risk of cord seroma after laparoscopic inguinal hernia repair with statistical significance.

P161

Communicating statistical methods, Simulation studies

Simulating probability of stopping to compare dose expansion design options

Rachel Pearson[1]

[1] Janssen

INTRODUCTION: In many phase 1 dose-finding trials, once the dose-escalation portion has identified one or several 'Recommended Phase 2 Doses' (RP2Ds), larger dose expansion cohorts at the RP2Ds are opened to gather more information on safety, efficacy and pharmacokinetics. In these dose expansion cohorts, toxicity should continue to be formally monitored, with many dose-limiting toxicity (DLT) monitoring rules to choose from. To aid selection of a rule for dose expansion, for each DLT rule under consideration, we produced a plot of the probability of stopping for toxicity over a range of 'true' DLT rates around the study target DLT rate. The aim was to identify rules with a low probability of stopping given a lower than target DLT rate and a high probability of stopping given a higher than target DLT rate.
APPROACH: In our study, we needed 25 participants in the dose expansion cohort to ensure sufficient power under our efficacy assumptions and we prespecified that we would monitor DLT rates after every 5 participants enrolled. Our target DLT rate was 30% and, using R, we simulated using 'true' DLT rates from 15% to 60%. For each 'true' DLT rate, we produced 10,000 binomial simulations for the full cohort (n = 25) and derived the cumulative number of DLTs at 5, 10, 15, 20, and 25 participants. Next, we applied the DLT monitoring rules to the simulated data to create indicators of whether the trial would be stopped due to toxicity under each of the DLT monitoring rules. These indicators were then summed across the simulations, by 'true' DLT rate, to produce the probabilities of stopping. Finally, we produced plots displaying the probability of stopping early for toxicity against the 'true' DLT rates for each DLT monitoring rule. The optimal rule would produce a sigmoid curve which rises steeply from very low probabilities to very high probabilities once the 'true' DLT rate crosses the study target DLT rate.
DISCUSSION: Visualizations of simulated probabilities of stopping for toxicity across different 'true' DLT rates are straightforward to implement and can support cross-functional conversations on DLT monitoring rule selection for dose expansion. When implementing DLT monitoring rules with cut-off parameters, such as Bayesian posterior probability rules, this approach can also be used to identify appropriate cut-off values.

Poster Presentations

P162

Simulation studies, Missing data

Simulating the Impact of Intercurrent Events and Missing Data for Clinical Trials

Sunita Rehal[1], Thomas Drury[1]

[1] *GlaxoSmithKline, United Kingdom*

The estimands framework [ICH E9(R1)] makes it clear any treatment effects targeted should include information about events that occur post-baseline (termed intercurrent events [ICE]) and how these will be handled. The impact of any ICEs will depend on the handling strategy chosen, and it is critical to assess the potential impact of ICEs as well missing data on the effects targeted at the design stage of a trial. Simulation under a range of assumptions is the clearest way to investigate this but is not always straightforward.
We focus on clinical trials where the goal is to describe the trends within patient changes in the response over time. In the case of repeated measures, patient outcomes at a future visit depend on previous patient responses. It is important to account for this conditional dependency by ensuring the outcomes are correlated between the repeated measures within the simulation before imposing a missingness mechanism to generate the missing values. This can be challenging for cases where there is a long sequence of follow up visits.
We present a generic simulation approach to generate dependent multivariate outcomes using Copula methodology. Using Copula is a simplistic approach that allows data generation for any distribution in a few simple steps by forming joint distributions based on a given dependence structure. Once a selection model has been applied to select subjects who have an ICE and when the ICE occurs, we create missing data for a monotone missing data pattern and then apply a withdrawal model to select subjects who withdraw from the study. We illustrate how these methods can be implemented via a simulation study and show how to check the correct distributions and dependence has been created.

P163

Simulation studies, Other (if none of the topics are applicable)

Simulation study of within-subject variability estimation: linear mixed effects & variogram analyses

Simon Baldwin[1]

[1] *University of Birmingham, United Kingdom*

CONTEXT: Longitudinal Biological Variability Studies (BVS) are crucial to estimate biomarker variability over time; ideally subjects attend visits with repeated measures taken multiple times per visit. Measurement error ('noise') is the key within-subject variability component estimated in BVS to evaluate biomarker repeatability, defined as the difference between an 'imperfect' and 'perfect' test (where multiple measurements on the same subject taken at almost the same time would be equal). Comparing true change ('signal') and measurement error within-subject variability estimates is one way biomarker repeatability can be evaluated; but prospective longitudinal BVS aren't always feasible and multiple measurements within-time-points within-subjects aren't always available. Methods to estimate 'signal' and 'noise' in different scenarios of repeated measures data is an important research question.
AIM: Issue guidance on 'signal' and 'noise' estimation when using linear mixed effects (LME) and [semi-]variogram analyses relative to different scenarios of repeated measures data.
METHODS: Simulated scenarios of repeated measures data will be generated of the form $Y_i(t) = \hat{\alpha} + \beta(t) + \alpha_i + \beta_i(t) + \omega_i(t)$; i denotes the subject $\{1 \le i \le n1\}$ and t denotes the visit time-point $\{0 \le t \le n2\}$ giving repeated measurements at n2 visits for n1 subjects. 'Change-from-baseline' and 'change-from-mean' models will be fitted to the data by LME and [semi-]variogram analyses. $\text{var}(\beta_i(t))$ and $\text{var}(\omega_i(t))$ are referred to as 'signal' and 'noise' respectively. Different subject numbers, visit time-points, non-attendance levels and visit time irregularities will be considered. Variation in simulation input parameters will also be considered based on two previous case-studies.
RESULTS: LME and [semi-]variogram analyses of 1900 repeated simulations per data scenario will be compared to input parameters.
CONCLUSION: Results will inform how guidance on 'signal' and 'noise' estimation when using LME and [semi-]variogram analyses can be issued relative to different scenarios of repeated measures data.

P165

Efficient clinical trial designs

Simultaneous evaluation of superiority and non-inferiority for sensitivity and specificity

Miyui Hara[1], Satoru Shinoda[1], Yusuke Saigusa[1], Kouji Yamamoto[1]

[1] *Yokohama City University School of Medicine, Japan*

In medicine, diagnostic test in medical fields is crucial when the gold standard methods are invasively or costly. Let a study consider a subject undergoes two simple tests, and whether the person is actually affected with the disease is diagnosed by the gold standard method. When comparing the diagnostic performance of two simple screening tests between a novel and an existing test, we often assess them by following, that is, whether the sensitivity or specificity of one test is superior to that of the other test.
Ideally, both the sensitivity and the specificity of a novel test should be superior to those of an existing test. However, as these measures have trade-offs, it is difficult to achieve superiority on both measures. Therefore, it is meaningful to evaluate sensitivity and specificity simultaneously to avoid making the unreasonable conclusion of superiority or inferiority of diagnostic performance based solely on one measure while completely ignoring the other.
This study proposes two statistical methods in the framework where we show the efficacy of a novel test only when 1) a novel test shows superiority to an existing test in at least one in sensitivity or specificity, and 2) it also shows non-inferiority to an existing test for the other measure. We evaluate the performance of the proposed methods in terms of type I error rate and power using Monte Carlo simulations.
The results of the simulation studies indicated that the proposed methods could adequately control the type I error rate. Furthermore, an example of application to the real data also shows the usefulness of the proposed methods.

P166

Machine learning methods for health, Biomarker discovery

Sparse group penalties for bi-level variable selection

Gregor Buch[1], Andreas Schulz[1], Irene Schmidtmann[1], Konstantin Strauch[1], Philipp S. Wild[1]

[1] *University Medical Center of the Johannes Gutenberg University Mainz, Germany*

INTRODUCTION: An important characteristic of many omics data sets is their intrinsic group structure due to high correlations or contextual similarities of features. Bi-level selection methods account for such groupings in the selection process to identify relevant variable groups and highlight their predictive members. One of the best known approaches of this kind combines the absolute shrinkage and selection operator (LASSO) with the group LASSO in an additive manner: sparse group LASSO (SGL). Since LASSO has some shortcomings that can be addressed by using alternative penalties, a generalization of SGL that enables combinations of other shrinkage terms is desirable.
METHODS: We propose a framework for sparse group penalties (SGP) that allows the combination of different SGL-style shrinkage conditions. Within this framework, we have combined the minimax concave penalty (MCP), the smoothly clipped absolute deviation (SCAD), the exponential penalty (EP) and their group versions analogous to SGL: sparse group MCP (SGM), sparse group SCAD (SGS) and sparse group EP (SGE). Corresponding objective functions were solved using the locally approximated coordinate descent, which we implemented in C++. The performance of the new methods in variable and group selection was compared with other bi-level selection methods (group exponential LASSO, composite MCP and group Bridge) in simulation studies.
RESULTS: SGE demonstrated superiority for variable and group selection in almost all settings where the number of observations exceeded the number of variables. In cases where there were fewer observations than variables, SGE was the best method when few groups contained predictive signals, SGM when a moderate amount of groups were relevant, and SGS when many groups were predictive. The classical SGL was always inferior to the other bi-level selection techniques in terms of variable and group selection, but its predictive performance was convincing in some situations.
CONCLUSIONS: Replacing the LASSO components in SGL with other penalties offers advantages with respect to several performance criteria, making approaches such as SGM, SGS, and SGE advisable over SGL. The benefits of these novel techniques are underlined by their ability to achieve better results than alternative bi-level selection methods, which SGL fails to do.

Poster Presentations

P168

Simulation studies, Missing data

Statistical methods of handling ordinal longitudinal response with intermittent missingness

Omololu Aluko[1], Henry Mwambi[2]

[1] University of the Free State, South Africa, [2] University of KwaZulu-Natal, South Africa

The rate of survival of human immunodeficiency virus (HIV) positive individuals resume to ameliorate with the consumption of highly antiretroviral therapy (HAART), but pulmonary disease prevalence has been growing unabated among them. The data was characterized with intermittent missing data due to patient's failure to disclose vital health information and absent on visit days. Handling missing data was a difficult challenge in the dataset. We analyzed the data under missing at random missingness assumption. We compared the effects of marginal and conditional models in the study. Amongst the methods, ordinal negative binomial model without any form of imputation performs greatly in simulation studies and real application than multiple imputation based generalized estimating equations (MI-GEE) and other models used.

P169

Efficient clinical trial designs

Statistical tests for seamless phase II/III design using short and long-term binary outcomes

Kenichi Takahashi[1], Ryota Ishii[2], Kazushi Maruo[2], Masahiko Gosho[2]

[1] MSD KK, Tokyo, Japan, [2] University of Tsukuba, Japan

The seamless phase II/III design combining phases II and III into a single trial has been shown growing interest for improving the efficiency of drug development, becoming the most frequent adaptive design type. It typically consists of two stages, namely an exploratory stage and a confirmatory stage, but the trial objectives being often different in each stage. The primary objectives are to select optimal experimental treatment group(s) in the first stage and compare the efficacy between the selected treatment and control groups in the second stage based on accrued data from both stages. We focus on a seamless phase II/III design, for which treatment selection is based on the short-term binary endpoint and treatment comparison is based on the long-term binary endpoint. For the situation in which the endpoints for treatment selection (short-term endpoint) and comparison (long-term endpoint) differ, we must consider the distribution of each endpoint and the correlation between them. We thus propose an exact conditional test for comparing the treatment effect between selected treatment and control groups in the final analysis, based on the bivariate binomial distribution and given the selected treatment with the most promising short-term endpoint response rate from an interim analysis. Additionally, we also propose the mid-p method based on the conditional exact distribution to improve conservativeness for an exact test. Simulation studies were performed to compare the type I error rate and the power of the combination test to our proposed methods for various scenarios. The proposed exact method controlled for type I error rate constantly at the nominal level, regardless of the number of initial treatments or the correlation between short- and long-term endpoints. In terms of the treatment comparison power, the proposed methods are more powerful than that based on the combination test in the scenarios, with only one treatment being effective.

P170

Efficient clinical trial designs

Superiority, Equivalence or Non-Inferiority? Selecting the Appropriate Trial Design

Nikki Totton[1], Steven Julious[1], Dyfrig Hughes[2], Jonathan Cook[3], Katie Biggs[1], Lizzie Coates[1], Andrew Cook[4], Catherine Hewitt[5], Simon Day[6]

[1] University of Sheffield, United Kingdom, [2] Bangor University, United Kingdom, [3] University of Oxford, United Kingdom, [4] University of Southampton, United Kingdom, [5] University of York, United Kingdom, [6] Clinical Trials Consulting & Training Ltd, United Kingdom

BACKGROUND: Randomised controlled trials can be designed with the objective to show superiority, equivalence or non-inferiority depending on whether their purpose is to demonstrate that a new health technology is better, the same as or no worse than the comparator. In some situations, it may not be obvious which is the most appropriate design to select for a given trial. In a survey completed in 2019 by the research team, 44% of respondents requested more guidance on how to select the most appropriate of these trial designs.
METHODS: In 2019, the team ran a consensus workshop for experts in the area of clinical trials considering a range of different stakeholders. To create a checklist of considerations when selecting the most appropriate trial design, the nominal group technique was implemented. This consisted of:
i) a preliminary round to elicit initial ideas,
ii) a discussion to clarify any of the items suggested and
iii) two voting rounds to decide on items for inclusion in the final checklist.
Consensus was demonstrated if 75% of participants voted either 'strongly (dis)agree' or '(dis)agree' with any item.
RESULTS: Fifteen experts attended the workshop and after the initial ideas round, twenty-eight discrete items had been identified which were categorised into six overarching categories. These categories relate to the: population, intervention, comparator, outcomes, feasibility and perspectives which were all deemed important considerations when selecting the appropriate trial design.
After the voting rounds, sixteen items were selected for inclusion and a further three items were added after discussion of the initial results with the oversight committee. A checklist of each of the nineteen items has been created with additional information, along with case study examples to demonstrate the use of the checklist.
CONCLUSIONS: To facilitate study team discussions on the most appropriate trial design we propose the nineteen-item checklist which will help to ensure all important elements have been considered.

P171

Ageing

Survival of Danish twins born 1870-2000

Philip Hougaard[1], Jacob v.B. Hjelmborg[2]

[1] Lundbeck, Denmark, [2] University of Southern Denmark, Odense, Denmark

Hougaard, Harvald and Holm (1992, Journal of the American Statistical Association, 87, 17-24) used frailty models to consider the survival of same-sex Danish twins born between 1881-1930 with follow-up until 1980 for twins, where both were alive at age 15. That dataset included around 9000 pairs of twins. This presentation gives an update to that analysis. For the birth cohorts 1870-1930, same-sex twins, where both were alive at age 6, are considered. For the birth cohorts 1931-2000, all twins are included. Follow-up is to 2022. This dataset includes around 50000 pairs of twins. The updated dataset has several advantages compared to the old. It is larger in several dimensions (such as earlier as well as later cohorts, longer follow-up, including also opposite-sex pairs and starting observation earlier in life). This, of course, allows for better precision but probably more interesting, it allows for studying time-trends. Furthermore, it is possible to discuss the appropriateness of shared frailty models for studying this problem.

Poster Presentations

P172

Efficient clinical trial designs, Simulation studies

SW-CRTs with binary outcome: Simulation study and application to a quality improvement trial

Christine Adrion[1], Ulrich Mansmann[1]

[1] *Ludwig-Maximilians-Universität (LMU), Institute for Medical Information Processing, Biometry and Epidemiology (IBE), Munich, Germany*

Due to several centre-related barriers to recruitment, stepped wedge cluster randomised trials (SW-CRTs) often modify their planned design and trial duration to reach the target sample size, which sometimes leads to an adaptation of the analysis strategy. We focus on a SW-CRT with a binary primary outcome, with participants being recruited and measured continuously over time but exposed rather briefly. We illustrate the statistical challenges using as case study a completed German pragmatic multicentre SW-CRT with a quality improvement intervention in the ICU setting involving a moderate number of 12 clusters (NCT03671447). As analytical approach for principal analysis, we specified a generalized linear mixed-effects model allowing the use of continuous time parameterizations and making appropriate adjustment for a secular time trend and exposure time. To address recruitment and implementation challenges, the recruitment duration was extended, while postponing pre-specified crossover dates and lengthening the post-rollout period. Several issues with the adherence to the randomisation schedule occurred (e.g. late or no implementation of the intervention in several clusters; delays in cluster recruitment). Additionally, imbalances concerning cluster sizes emerged.
Based on this motivating example, we conduct a simulation study on allocation characteristics to explore the statistical impact of cluster-level deviations from planned design features on the attained power and performance measures regarding the intervention effect estimate. Simulations are run in R using the lme4 package.
In our simulated SWT scenarios, mixed-effects model results were sensitive to deviations from model assumptions and planned design features (e.g. regarding time confounding the intervention effect). Simulation studies are essential tools in understanding the impact of modifications of the design and model misspecifications in a SW-CRT with a complex configuration. Results provided by these simulated scenarios enable to assess the robustness of the chosen analysis strategy in a real SW-CRT to ensure that the intervention effect estimator maintains the nominal 5% significance level and is also reasonable unbiased. Structured reporting on how they were planned and performed is necessary to make their influence on model choice and supplement sensitivity analyses transparent.

P173

Miscellaneous

Causal Pathways from Parental Supply to Alcohol-related harm – A Causal Mediation Analysis using TMLE

Philip Clare[1], Wing See Yuen[2], Alexandra Aiken[2], Amy Peacock[2], Raimondo Bruno[3], Louisa Degenhardt[2], John Horwood[4], Delyse Hutchinson[5], Kypros Kypri[6], Richard Mattick[2], Nyanda McBride[7], Jim McCambridge[8], Jackob Najman[9], Monika Wadolowski[10], Tim Slade[11]

[1] *Prevention Research Collaboration, University of Sydney,* [2] *National Drug and Alcohol Research Centre, UNSW Sydney,* [3] *University of Tasmania, Australia,* [4] *University of Otago, New Zealand,* [5] *Deakin University, Australia,* [6] *University of Newcastle, Australia,* [7] *Curtin University, Australia,* [8] *University of York, United Kingdom,* [9] *University of Queensland, Australia,* [10] *Macquarie University, Australia,* [11] *University of Sydney, Australia*

INTRODUCTION: Recent research suggests parental supply of alcohol (PSA) is associated with higher risk of subsequent alcohol-related harm. However, the mechanisms connecting PSA and harm remain unclear. Due to the complexity of longitudinal data, analysis of such overall effects using standard analytic methods can be biased. This study used causal mediation analysis to consider potential mediators of the PSA-harms relationship using robust methods to minimise risk of bias.
METHODS: Using data from the Australian Parental Supply of Alcohol Longitudinal Study (APSALS; n=1906), we examined the mediating effects of family (parental monitoring, alcohol-specific rules, and alcohol norms) and peer (peer supply of alcohol, peer substance use and peer disapproval of substance use) on the associations between PSA and binge drinking, alcohol-related harms, and alcohol use disorders. To reduce risk of bias, analyses used targeted maximum likelihood estimation using machine learning.
RESULTS: Evidence showed indirect effects of PSA through both family and peer variables on all outcomes, with 69% of the effect of PSA on monthly binge drinking explained by natural indirect effects, and 63% of the effect of PSA on experience of alcohol-related harm.
CONCLUSIONS: There is evidence that for mediation between PSA and subsequent harm through both family and peer variables, with around two thirds of the association explained by natural indirect effects through these mediators. This suggests further potential avenues for intervention to reduce the risk of harm to adolescents.

P174

Miscellaneous

The effect of sample size on selecting a satisfactory descriptive model with the MFP approach

Willi Sauerbrei[1], Patrick Royston[2], Aris Perperoglou[3]

[1] *Medical Center - University of Freiburg, Germany,* [2] *MRC Clinical Trials Unit at UCL, London, United Kingdom,* [3] *School of Mathematics, Statistics and Astrophysics, Newcastle University, United Kingdom*

Data analysts are often faced with many potential predictors with a requirement for a satisfactory descriptive model. The task involves the selection of a subset of variables with a relevant influence on the outcome. With continuous variables a suitable function needs to be determined.
The multivariable fractional polynomial (MFP) approach combines selection of variables using backward elimination (BE) with the fractional polynomial function selection procedure (FSP) for continuous variables (https://mfp.imbi.uni-freiburg.de/). The results from BE and the FSP depend on testing and hence on the sample size and stopping criteria (e.g. p-values, AIC, BIC). Provided the sample size is 'sufficient', MFP is a suitable approach for the selection of a descriptive model, which aims to capture the data structure parsimoniously (in contrast to a predictive model).
Ideally, an MFP model fits the data well, and is simple, interpretable, transferable and stable. The latter means that it hardly changes in similar data. In data with a 'large' sample size, MFP probably identifies all relevant variables and selects satisfactory functional relationships for continuous variables (e.g., log or non-monotonic function). However, what is a 'large' sample size in the context of a multivariable model with about 10 to 30 variables, some of which are continuous? Can we derive an acceptable MFP model with 500 observations or do we need 10 000, or many more?
In examples with normal-errors, binary and Cox regression, we use two large datasets (about 20 000 observations) and a big data set (> 100 000 events in >400 000 patients) to investigate the effect of sample size on the model selected in subsamples. As a start for investigating model stability in smaller datasets, we compare the results from three non-overlapping replications with the same sample size.
Considering the selected model from all data as the 'gold standard', we show that the same or very similar models are selected in much smaller data sets and we argue that MFP with BIC as the selection criterion yields satisfactory descriptive models in data with a 'larger' sample size. We tentatively suggest guidance on the sample size needed in various practical cases.

P175

Communicating statistical methods

The Evaluation of Repeated Events in the PLEASANT dataset

Kelly Grant[1], Steven Julious[2]

[1] *University of East Anglia, United Kingdom,* [2] *University of Sheffield, United Kingdom*

Asthma is a common health condition affecting school-aged children. Research suggests fewer asthma prescriptions are collected in summer, compared to the rest of the year, and more medical appointments for asthma difficulties occur in September. The PLEASANT study investigated whether sending a medication reminder letter in July, to parents or guardians of school-aged children with asthma, would increase prescriptions uptake in August, and decrease unscheduled medical appointments in September.
The study was conducted in GP practices across England and Wales, as a cluster randomised trial. Standard analysis methods were used – binary logistic regression, negative binomial regression, Cox proportional hazards – allowing for the effects of clustering. The study found evidence of an increase in prescriptions uptake in August for the intervention group, however, no evidence of decreasing unscheduled contacts at the primary time point of September. Although, there was some evidence of effect over a wider time interval.
An issue with the analyses undertaken with the PLEASANT trial is they neither took account of time between medical contacts (negative binomial), nor recurrent events (logistic and Cox regression).
A re-analysis of the PLEASANT trial investigates whether accounting for repeated (recurrent) events per participant, impacts trial conclusions. The conditional frailty model is used, with the frailty term at GP practice level to account for clustering. Within participant event dependence and an increasing risk of subsequent events are also accounted for. A rare events bias adjustment is applied if few participants have recurrent events, and truncation of small event risk sets is explored as sensitivity analyses, to improve model accuracy.
Overall, the conditional frailty model results are consistent with the negative binomial model PLEASANT results in terms of effect size, matching original conclusions, though with greater precision – as assessed by more narrow confidence intervals.
Recurrent event survival analysis methods are recommended when there are recurrent events for a study outcome. These analysis methods potentially increase statistical power by including extra information on time, leading to higher precision compared to standard methods.

Poster Presentations

P176

Meta-analysis

The Hierarchical Bias-Corrected Meta-Analysis Model

Pablo Emilio Verde[1]

[1] *Coordination Center for Clinical Trials, Heinrich Heine University Düsseldorf, Germany*

During the COVID-19 pandemic, we have faced the need to make meta-analyses including a mixture of all the available evidence, namely observational studies with varying quality. The main issue when combining disparate evidence in a meta-analysis is that we are not only combining results of interest, but we are also combining multiple biases. Therefore, commonly applied meta-analysis methods may lead to misleading conclusions.
In this work, we present a new Bayesian hierarchical model, called the Hierarchical Bias-Corrected (HBC) meta-analysis model, to combine different study types in meta-analysis.
This model extends the Bias-Corrected (BC) meta-analysis model (Verde 2021) when each study reports multiple risk factors, but not necessarily the same ones in each study.
The BC is based on a mixture of two random effects distributions, where the first component corresponds to the model of interest and the second component to the hidden bias structure. In this way, the resulting model of interest is adjusted by the internal validity bias of the studies included in a systematic review. The HBC allows exchangeability between the bias correction and the model of interest when different risk factors are considered simultaneously.
We illustrate the HBC model with a real meta-analysis, which investigates the influence of baseline risk factors on complications and mortality in patients with positive COVID-19. We show that the use of simple random-effects meta-analysis models exaggerates the risk of comorbidities in those patients. The HBC model has been implemented in the R package jarbes (Verde 2022), which facilitates its application in practice.

P177

Simulation studies

The impact of fitting linear mixed models to ordinal outcomes

Petra Baumann[1]

[1] *Medical University of Graz, Austria*

In many disciplines, it is common to fit linear models to ordinal outcomes (OO) and this is often sanctified by claiming that the OO is a crude representation of a latent continuous and normally distributed outcome with equidistant categories. While in reality these claims are impossible to prove and research suggests that they are unlikely to hold, we created this "ideal world" in a Monte Carlo simulation to assess the potential risk of this practice.
We created equidistant OOs (with 3-11 categories) by cutting up normally distributed continuous variables. We compared results from two-level random intercept linear vs. ordinal models with one continuous and one rank predictor on the lower level and one continuous predictor on the upper level. To assess ICC estimation, we used null models without predictors. Sample size was 20.000 (100 groups with 200 observations), number of iterations was 1.000 for each condition. Model estimates cannot be compared directly, but variability of estimates across simulations can be compared to the estimated standard errors of the estimates ("optimism"). Another measure that is comparable across different model types is the ICC. We also compared estimates from linear mixed models with ordinal outcomes (LMM_OO) vs. baseline continuous outcomes by calculating relative bias and the coverage rate of the 95% confidence interval for fixed effects.
Regarding optimism, we found that LMM_OOs are overconfident with the continuous predictor but perform well for the effect of the rank predictor. For the effect of the upper-level predictor, the LMM_OOs seem to perform better. But optimism is only a measure of the consistency of the estimation in light of the estimated standard errors, and not of bias. The coverage rates of the 95% C.I. and the relative bias of the fixed effects revealed that estimation is severely biased in LMM_OOs. Moreover, the size of the bias seemed to be related to the number of categories in the outcome and "jumped" from under- to overestimation at a seemingly random point. At least the direction of the effect was estimated correctly in most cases, but the ICC was always underestimated in linear LMM_OOs. Based on these results, we discourage the use of LMM_OOs.

P178

Miscellaneous

The number of cases, mortality and treatments of viral hemorrhagic fevers: a systematic review

Drifa Belhadi[1], Majda El Baied[2], Guillaume Mulier[2], Denis Malvy[3], France Mentré[1], Cédric Laouénan[1]

[1] *IAME, INSERM, UMR1137, Université Paris Cité, France,* [2] *AP-HP, Paris, France,* [3] *Inserm 1219, University of Bordeaux, France*

BACKGROUND: Viral hemorrhagic fevers (VHFs) are a group of diseases characterized by outbreaks occurring irregularly and hard to predict. New treatments need to be evaluated but due to specificities of VHF, difficulties arise when conducting clinical studies.
OBJECTIVES: To assess the state-of-the-art on 18 VHFs, we systematically reviewed the reports (WHO, CDC) and literature (MEDLINE, Embase, CENTRAL) that included specific results on cases number, mortality and treatments. The final objective is to use these results to conduct a Bayesian Decision Analysis (BDA; Isakov et al. J Econom 2019) to evaluate the optimal sample sizes and type I errors for future clinical trials.
METHODS: The search was conducted in January 2020 based on PRISMA guidelines (PROSPERO CRD42020167306). Following the study selection process, qualitative and quantitative data were extracted from articles. A narrative synthesis approach by VHF was used for each outcome. Descriptive statistics were conducted including world maps of cases number and case fatality rates (CFR); summary tables by VHF, country, time period and treatments studies.
RESULTS: We identified 141 WHO/CDC reports and 126 articles meeting the inclusion criteria. Most of the studies focused on 7 VHFs, with 90 studies focusing on Ebola Virus Disease (EVD). EVD outbreaks were reported in Africa, where Sierra Leone reported the highest cases number (14,124 cases; CFR=28%), mainly due to the 2014/16 outbreak. 17 EVD treatments studies were identified, among which a recent randomized trial reported 2 treatments (Mab114 and REGN-EB3) as significantly reducing EVD mortality. Crimean-Congo hemorrhagic fever (CCHF) in 31 studies were reported in Africa, Asia and Europe, where Turkey reported the highest cases number (6,538 cases; CFR=5%) and Afghanistan the last outbreak in 2016/18 (293 cases, CFR=43%). 9 CCHF treatments studies were identified, mainly assessing ribavirin with inconsistent efficacy results.
CONCLUSIONS: We found that most studies focused on a restricted number of VHFs, mainly EVD. EVD cases were prominent in Africa and several studies were conducted to find a treatment. There was a lower number of studies on CCHF with cases reported in Africa, Asia and Europe. However, no strong evidence of CCHF treatment was identified. All those results will be used in the BDA to inform the design of future VHF trials.

P179

Prediction models

Time-varying frailty multistate model in the presence of time-dependent covariates

Robabeh Ghodssi-Ghassemabadi[1], Ebrahim Hajizadeh[1], Mahmoud Mahmoodi[2]

[1] *Tarbiat Modares University, Tehran, Iran,* [2] *Tehran University of Medical Sciences, Iran*

Multivariate survival data have an important role in clinical studies. These data arise from studies in which subjects are clustered into groups or each subject experiences several events of different (the same) types during the study. Multi-state model is one of the usual methods for modeling multivariate survival data. Predictions in survival analysis become adjustable when considering intermediate events, so the result of multistate survival models are more accurate than considering just one event.
Homogenous multistate models have been widely used for progression of diseases and health conditions. However, subjects are not likely to share the same natural history of disease or health condition despite having the same prognostic factors. Incorporation of frailty into the multi-state model is a way of identifying factors affecting disease history while quantifying this heterogeneity among subject, while ignoring this individual heterogeneity might cause biased estimate of factors affecting transition intensities. One of the key assumptions in frailty models is that the individuals have a constant frailty during the entire disease process, while this might be a restrictive assumption.
We proposed a time-varying frailty multistate model with time dependent covariates in which the frailty varies over time intervals using a power parameter approach [1]. We conducted extensive simulation studies to assess the robustness and performances of the proposed model in comparison to a constant frailty multistate model with time-dependent covariates. As an application, we implemented the proposed model to data from colorectal cancer patients to investigate clinicopathological factors affecting disease progression over time.

Poster Presentations

P180

Personalized medicine

Tree-based methods to identify subgroups and personalized effects of medication adherence on costs

Johannes Wendl[1], Alexander Hapfelmeier[1]

[1] *Institute of General Practice and Health Services Research at Technical University of Munich, Germany*

Adherence to medication of chronically ill patients is defined as the extent to which a patient follows agreed recommendations from a health care provider. It is crucial for treatment success and hence decreased future health costs. However, commonly used regression models failed to show this relation. We therefore aimed to distinguish subgroups with different effects and patients with individual effects.
We used German claims data collected between 2012 and 2015 to define a large cohort of Diabetes Type 2 patients (n=85,142) and estimated the overall associations between adherence and costs with linear regression models. We extended our analysis using model-based trees to identify subgroups of patients by sociodemographic and health status variables with different estimated adherence effects. Moreover, we applied model-based random forests to estimate patient individual adherence effects. To measure performance of our model we developed a new precision based approach. Therefore we compared the personalized effects predicted by the forest for some training data to the effects obtained when the regression models of the forest are refit to the test data, conditional on the fixed structure of the forest. A regression of these two estimates with 95% prediction intervals was used to identify patients where we can expect a negative adherence effect with the given certainty based on the respective prediction of the forest.
While a simple linear regression model showed positive association between adherence and costs overall, both, model-based trees and forests identified patients with negative effects. Forest specification with smaller terminal node size resulted in higher variance of estimated adherence effects but lower precision, indicating a respective trade-off.
Our approach shows that tree-based models can identify patients with different effects. With our new approach we can also measure the performance of such models through precision of personalized effects. Identified patients can form target groups for adherence-promotion interventions with the aim to increase health and decrease associated costs.

P181

Machine learning methods for health

Type 2 diabetics at highest risk for heart failure: a survival model based on EHRs

Ilaria Gandin[1], Sebastiano Saccani[2], Andrea Coser[2], Scagnetto Arjuna[3], Andrea Di Lenarda[3], Giulia Barbati[1]

[1] *University of Trieste, Italy,* [2] *Aindo, Trieste, Italy,* [3] *ASUGI, Trieste, Italy*

Type 2 diabetes mellitus (DM) is a global epidemic and is expected to affect over 592 million people worldwide by 2035 (Guariguata 2014). Recent studies suggest that patients with DM have over twice the risk of developing heart failure (HF) than patients without diabetes mellitus (Dei Cas 2015). The present study is aimed to build a survival model that takes in account a large and heterogeneous set of clinical factors and investigate the risk of developing HF in diabetic individuals.
The study cohort is formed by patients of the Cardiovascular Observatory of Trieste, Italy (Gasperoni 2018). Information consists of Electronic Health Records (EHRs) extracted from clinical and administrative data obtained as part of routine medical care. It includes a variety of variables: diagnostic codes, hospitalizations, laboratory tests, procedures, cardiovascular drugs prescriptions, comorbidities. The primary endpoint is the onset of HF. Time-to-event survival analysis is carried out on baseline individual's characteristics with two approaches. First, a multivariable Cox proportional hazards (CPH) regression model is estimated. In the second approach, we estimate a model based on deep learning techniques (DeepSurv, Katzman 2018) in which a neural network is used to represent the hazard function, working as a nonlinear generalization of CPH models. Given the high number of predictors, we apply a variable selection process involving the one-by-one analysis of variables impact.
The study involves 10,614 individuals. During a median (IQR) follow-up of 65 (68) months, we detect 1840 (17.3%) failure events. In our analysis we observe an increase in the performance in the deep learning approach. Using the CPH model we obtain a c-index of 0.730 (SD 0.007), whereas using the DeepSurv model the c-index increase up to 0.762 (SD 0.009). Among the top-impact predictors we find age, specific ECG parameters, diuretic drugs, hemoglobin, systolic blood pressure.
Our preliminary results suggest that deep learning techniques provide more flexibility and could be an effective approach for survival analysis in the context of EHR.

P182

Meta-analysis

Using extracted data from Kaplan-Meier curves to compare COVID-19 vaccine efficacy

Amarit Tansawet[1], Songporn Oranratnachai[2], Pawin Numthavaj[3], Ammarin Thakkinstian[3]

[1] *Faculty of Medicine Ramathibodi Hospital, Bangkok, Thailand,* [2] *Sripat Medical Center, Chiang Mai University, Thailand ,* [3] *Mahidol University, Bangkok, Thailand*

BACKGROUND: Although various COVID-19 vaccines have shown efficacy against placebo in randomized clinical trials, no head-to-head comparisons are yet available. This study aims to compare the efficacy of available COVID-19 vaccines.
METHODS: Vaccine trials searched in May 2021 were included. Data were extracted from Kaplan-Meier (KM) curves using the WebPlotDigitizer program for the individual participant (IP) data simulation. A mixed-effect acceleration failure model with log-logistic and Weibull distributions was used to estimate relative effects for individual vaccines as well as grouped by class: inactivated virus, mRNA, and viral vector. Primary studies were considered as the random effect in the model. Hazard ratios (HR) were estimated and compared across vaccine groups.
RESULTS: All vaccines were efficacious in lowering symptomatic infection compared to placebo. CoronaVac, Ad26.COV2.S, ChAdOx1 nCoV-19, rAd26/rAd5, WIV04, HB02, and BNT162b2 showed 7.61 (4.50, 12.87), 6.77 (4.08, 11.24), 5.01 (2.93, 8.57), 4.50 (2.52, 8.01), 3.90 (2.04, 7.45), 3.18 (1.62, 6.21), and 2.15 (1.22, 3.78) times significantly higher risk of infection than mRNA-1273. mRNA vaccines were the most efficacious vaccine group compared to inactivated virus and viral vectors with HRs (95% CI) of 0.27 (0.20, 0.37) and 0.28 (0.21, 0.37), respectively.
CONCLUSIONS: Although all vaccines showed significant protection compared to no vaccination. mRNA vaccines, including mRNA-1273 and BNT162b2, showed the highest efficacy in preventing symptomatic COVID-19 infection. Simulated IP data from KM curve might allow treatment comparison when there is no primary study comparing between active treatments.

P183

Prediction models, Medical device clinical studies

Validation and agreement of clinical prediction models for diagnosis of COVID-19

Nadia Dardenne[1], Médéa Locquet[2], Anh Diep[1], Allison Gilbert[3], Sophie Delrez[3], Charlotte Beaudart[2], Christian Brabant[2], Alexandre Ghuysen[3], Olivier Bruyère[2], Anne-Françoise Donneau[1]

[1] *Biostatistics Unit, University of Liege, Belgium,* [2] *Epidemiology and Health, Economics, University of Liege, Belgium,* [3] *Emergency Department, University Hospital Center, Liege, Belgium*

BACKGROUND: Different prediction models have been developed to help clinicians in rapid triage of high-risk of Covid-19 patients to avoid hospitals overcrowding. The objectives of our research were to review the latest prediction models, to evaluate and compare them using an independent dataset incorporating several waves of the pandemic.
METHOD: The literature review was conducted during the first quarter 2020 using Medline (via Ovid) and Scopus databases. Based on rigorous inclusion and exclusion criteria, 13 articles were finally considered. Among these 13 articles, 6could have been reconstructed and assessed using diagnostic tests, like sensitivity (Se) and negative predictive value (NPV), discrimination (Area Under the Roc Curve (AUROC)) and calibration measures. Agreement between models using the Kappa's coefficient and the Intraclass Correlation Coefficient (ICC) was also examined. The Fully Conditional Specification method of the multiple imputations approach was applied to deal with missing values. All statistics were realised on a total of 1618 adult patients from two Emergency Department triage centres in Liège, Belgium, during the first two waves of the pandemic. A sensitivity analysis was also conducted by wave.
RESULTS: Either due to missing information (>20%) or the impossibility to reconstruct the model (lack of information, country-specific variables, employed methodologies), 7 model were discarded. The 6 selected models successfully discriminated patients in their original studies. When comparing prediction models using the independent dataset, models based on symptoms and/or risk exposure alone were less efficient (AUROC < 0.80) than those including biological parameters and/or radiological examination; however all of them showed good calibration. Good results were also found for Se and NPV with values > 0.75. Finally, poor agreement (Kappa and ICC <0.50) were observed. The derived results remained identical in both waves of patients.
CONCLUSION: Discrimination and calibration measures highlighted the importance of radiological examination to obtain more efficient models and revealed the difficulties to find an easy-to-use tool to help clinicians to classify patients at risk of COVID-19. Moreover, as the pandemic has continued to evolve with new scientific and medical advances, carrying out a living systematic review would be recommended.

Poster Presentations

P184

Machine learning methods for health, Communicating statistical methods

Variable Selection and Estimation Using Group Lasso for Ordinal Outcomes

Mojtaba Lotfaliany[1], Tara Johnson[1], Sarah Hosking[1], Lars Kayser[2], Natalie Hyde[1], Lesley Berk[1], Mohamed Abdelrazek[3], Mohammadreza Mohebbi[4]

[1] IMPACT – the Institute for Mental and Physical Health and Clinical Translation, Melbourne, Australia, [2] Department of Public Health University of Copenhagen Copenhagen, Denmark, [3] Software Engineering and Internet-of-Things, School of Information Technology Deakin University, Melbourne, Australia, [4] Biostatistics Unit, Faculty of Health, Deakin University, Melbourne, Australia

Despite its efficiency in variable selection in developing predictive models, the group lasso approach was rarely used in health-related research, particularly for ordinal outcomes. The lack of such studies can be attributed to an additional assumption needed for models when the outcome is ordinal: homogeneity of threshold-specific effects.
This study aims to identify the most important predictors of computer use and three-year change in computer use, in a large longitudinal study. Group lasso was selected for this study due to the presence of categorical predictors as well as high multi-collinearity among them. To develop models that provide a single effect estimate collapsed over thresholds, constrained continuation models was used by restructuring the data. The robustness, validity, reliability of the results was assessed through a range of sensitivity analyses including 10-fold cross-validation for each penalty, missing information imputation.
Models with no penalty terms showed a high number of predictors, a low level of objective term and expected generalization error. Increasing the penalty resulted in a reduction in the number of parameters in the model, reducing the variance, as well as an increase in bias and therefore objective function. Importantly, the expected generalization error of the models remained steady in low-to-moderate penalties and only increased in very high penalties.
Our results show that group lasso models can be used with high robustness, validity, reliability for identifying the most important predictors for ordinal outcomes where there are a high number of grouped predictors with high multi-collinearity among them. Applying the group lasso models revealed important predictors of computer use and its change over time.

P185

Medical device clinical studies

Video Observed Therapy Device Improves Probability of Tuberculosis Therapy Adherence

Dorcas Kareithi[1], Samuel Olufemi Amoo[2], Tijani Bosun[2,3], Tochukwu Onuigbo[2], Steven Karera[3], Joy Oraegbu[2], Salako Babatunde[2]

[1] Newcastle University, Newcastle upon Tyne, United Kingdom, [2] Nigeria Institute of Medical Research, Lagos, Nigeria, [3] Co-Creation Hub, Lagos, Nigeria

BACKGROUND: Directly observed therapy (DOT), the standard of care for monitoring patients on treatment for tuberculosis (TB), requires substantial health department resources and can be inconvenient and disruptive for patients. On the other hand, Video Observed Therapy (VOT) is a method of adherence monitoring where patients transmit digital images of their treatment intake to a central location for review; either synchronously or asynchronously. The use of VOT in some developed countries has been documented to be acceptable, cost-effective, and improve patient commitment to treatment. Considering this, VOT presents an option that can be explored for developing countries like Nigeria with a high TB burden, thereby bringing about a reduction in TB burden and ultimately its elimination.
METHODS: The study was a two-arm individually randomized clinical trial conducted by the Nigeria Institute of Medical Research (NIMR) in Lagos, Nigeria. 100 participants were recruited to this study and randomized into either treatment (VOT) or control (DOT). Data was collected through the NimCure mobile app and patient records. We compared treatment outcomes between VOT and DOT and assessed average treatment effects adjusting for gender and age.
RESULTS: A high rate of 100% adherence to TB treatment was achieved in both groups, (97% of VOT and 79% of DOT). More females than males achieved 100% adherence for VOT, while more males achieved 100% adherence in DOT. There was a varied distribution across all age groups. The average adherence, if all patients were to use VOT, would be 0.1650 more than the average that would be if the patients did not use VOT.
CONCLUSION: VOT through the NimCure smartphone app achieved the adherence objective, we, therefore, recommend that VOT can be used for remote monitoring and management of Tuberculosis patients. VOT, therefore, presents improved adherence to TB treatment. NimCure can therefore be recommended for use in low-income settings.

P186

Prediction models

Wavelet analysis of the COVID-19 pandemic

Kazumi Omata[1], Ayumi Shimazaki[1], Hiroaki Mitsuya[1],

[1] National Center for Global Health and Medicine, Tokyo, Japan

In an infectious disease epidemic, when the pathogen is a virus, mutations in the genes of the virus cause a wave of infection epidemics. For the COVID-19 pandemic, many variants (B.1.1, B.1.1.284, B.1.1.214, B.1.1.7, AY.29, BA.1.1.2, etc.) had triggered epidemic waves and six infection waves were observed worldwide by June 2022. Infection waves, like physical waves, have information such as wavelength, amplitude, and phase. In this study, we analyze time-series data on the number of reported COVID-19 positive individuals by performing a wavelet analysis and examine the phase of infection waves, in particular. Using time-series data by prefecture in Japan, by municipality in Tokyo, and by country in the world, we aim to discover the spatial and temporal characteristics of the COVID-19 pandemic. The wavelet analysis derives transient relationships between non-stationary time series data. Using this technique, various infectious-disease epidemics were examined, and we have applied it to the COVID-19 epidemics. In Japan's prefectural analysis based on the phase of the infection wave in Tokyo, the outbreak of infection waves tended to be slower than Tokyo in the surrounding prefectures (Saitama, Kanagawa, Chiba, and Gunma). This delay was also seen in the prefectures connecting prefectures which include large cities such as Osaka, Nagoya, and Fukuoka. The prefectures in the southwestern area of Japan also exhibited slower infection waves. Other prefectures showed no statistically significant difference from Tokyo. These results provide a picture of the spread of infection waves from large cities to surrounding areas, and suggest that in the case of infection control measures such as urban lockdowns, estimating such characteristics of infection waves may increase the effectiveness of countermeasures when implementing them. The arrangement of the transportation network and the topography of the country are thought to have a significant impact on the transmission of infection waves, and we discuss the relationship among them. The analysis of each municipality in Tokyo did not produce characteristic results, and in contrast, the analysis by country around the world revealed the unique spread of infection in each country. As for the latter, further analysis will be of interest with respect to characterizing the COVID-19 epidemics.

P187

Prediction models, Joint models for high dimensional data

When Cox's regression works well but could work better: time-to-event confidence interval adjustment

Lubomir Stepanek[1], Filip Habarta[2], Ivana Mala[2], Lubos Marek[2], Ladislav Stepanek[1]

[1] First Faculty of Medicine, Charles University, Prague, Czech Republic, [2] Faculty of Informatics and Statistics, Prague University of Economics and Business, Czech Republic

The Cox proportional hazard model could identify whether groups varying in duration of manifested COVID-19 differ in the expected persistence of COVID-19 antibodies seropositivity. An estimate of the overall expected duration of positive antibodies, i. e., between the disease beginning and antibody (serology) test, assumes that a negative antibody test result means a patient's antibodies got already under a test cut-off. However, if a patient has a negative antibody test, they might already get below the test cut-off or even not reach the cut-off (yet), considering their antibodies are still growing when tested. Those two situations cannot be distinguished without follow-up data and, thus, have to be addressed by a statistical adjustment.
Thus, the statistical problem is to estimate a proportion of those with negative antibody tests whose antibodies are still growing, i. e. those who would get positive antibody tests when tested later. Unlike in the group with a positive antibody test, the Cox regression, although correctly applied to such data, systematically underestimates the expected duration of the antibodies having in a group of individuals with antibodies measured as negative, which was mathematically proven in the study. Furthermore, varying the proportion's prior distribution (random uniform, normal, exponential) parameters of those with the negative antibody test but continuous antibody-producing (i. e., antibody-positive if tested later) helps to estimate how much both the overall expected persistence of positive antibodies having (seropositivity) and a proportion of seropositive patients in a given time point is underestimated. Formulas for appropriate confidence intervals of both the statistics depending on prior distributions were derived. We applied the theory to a dataset of about 150 patients' survival COVID-19 data and calculated adjusted confidence intervals for antibody-positive proportions in given time points. That could help, besides others, to precisely (!) estimate when it is possible to postpone a COVID-19 vaccination with a negative antibody test.
The proposed statistical adjustments of both confidence intervals seem to make Cox's regression outputs more precise in the given setting, which may be helpful even for the COVID-19 vaccination time management.

P188

Prediction models

Multinomial prediction models for methotrexate outcomes in rheumatoid arthritis

Celina Gehringer[1], Glen Martin[2], Kimme Hyrich[1,3], Suzanne Verstappen[3,4], Jamie Sergeant[1,4]

[1] Centre for Epidemiology Versus Arthritis, Division of Musculoskeletal and Dermatological Sciences, Manchester Academic Health Science Centre, University of Manchester, United Kingdom, [2] Centre for Health Informatics, Division of Informatics, Imaging and Data Sciences, University of Manchester, United Kingdom, [3] NIHR Manchester Biomedical Research Centre, Manchester NHS Foundation Trust, Manchester Academic Health Science Centre, United Kingdom, [4] Centre for Biostatistics, Manchester Academic Health Science Centre, University of Manchester, United Kingdom

INTRODUCTION: Methotrexate (MTX) is the recommended first line therapy in patients with rheumatoid arthritis (RA), but around 40% of patients do not respond to treatment after 6 months. Our previous systematic review and meta-analysis identified that clinical prediction models of MTX outcomes suffered from methodological limitations, including a lack of validation, suboptimal handling of missing data, and no considerations of competing risks, such as patients that discontinue due to adverse events (AEs). The objectives of this study were to (i) predict outcomes of MTX therapy at 6 months, in terms of low disease activity (LDA) and AEs, using a multinomial framework, (ii) update prognosis at 3 months (3m), (iii) internally validate these models to assess performance.
METHODS: A multinomial logistic regression (MLR) of outcomes 1) not in LDA, used as reference, 2) LDA, and 3) discontinuation due to AEs was developed using data from the national multi-centre Rheumatoid Arthritis Medication Study (RAMS). The updated model was conditional on patients not being in LDA at 3m. Missing data was handled using multiple imputation, and models were internally validated through bootstrapping. Calibration slope and intercept were assessed, and discrimination was measured using pairwise c-statistics.
RESULTS: A total of 1632 patients were included in the analysis. At 6 months, 756 (46%) patients achieved LDA, 730 (45%) were not in LDA, and 146 (9%) discontinued due to AEs. The updated model at 3m included 1179 (72%) patients as they had not achieved LDA. After bootstrapping, the baseline model showed good discrimination for outcomes 2vs1 with a pairwise c-statistic of 0.73 (0.71, 0.76), which was 0.71 (0.68, 0.74) at 3m. For outcomes 3vs1 this was 0.53 (0.48, 0.58) and 0.57 (0.51, 0.63), respectively. The calibration slope at baseline was 0.99 (0.86, 1.12)(2vs1) and 0.73 (0.24, 1.20)(3vs1), and at 3m 0.97 (0.78, 1.12)(2vs1) and 0.80 (0.31, 1.28)(3vs1).
CONCLUSION: Our MLR models handled outcomes of disease activity and AEs simultaneously. We were able to predict LDA (2vs1) more accurately than AEs (3vs1). This approach provides more clinically useful and realistic predictions of MTX outcomes, as competing risks such as AEs have previously been ignored.

MS1 Check what is missing with initial data analysis

Lara Lusa[1], Katherine Lee[2], Carsten Oliver Schmidt[3], Marianne Huebner[4]

[1] University of Primorska, Slovenia, [2] Murdoch Children's Research Institute, Australia, [3] Institute of Community Medicine, University Medicine of Greifswald, Germany, [4] Michigan State University, United States

Missing data are common in most research studies, and how to handle this should be specified a priori in a statistical analysis plan. Missing values are therefore also a major point of concern in Initial data analysis (IDA), which aims for transparency and integrity by providing reliable information about properties of a dataset to enable researchers to perform the statistical analyses in a responsible manner and correctly interpret the obtained results. IDA is a valuable tool that helps reveal the potential impact of missingness and evaluate the appropriateness of the intended choices to handle missing data. In this talk we provide a checklist for common tasks of IDA related to missing values at the unit and item level, investigate the types and patterns of missingness in a dataset using appropriate summaries and visualizations. The examples are based on a longitudinal dataset from the Survey of Health, Ageing and Retirement in Europe (SHARE) with medical, economic and social characteristics. IDA can also be used to guide the construction of sampling weights that take non-response into account.

TG1

Berkson measurement error and the regression calibration approach

Victor Kipnis[1], Lawrence Freedman[2], Pamela Shaw[3]

[1] National Cancer Institute, United States, [2] Biostatistics and Biomathematics Unit, Gertner Institute for Epidemiology and Health Policy Research, Tel Hashomer, Israel, [3] Division of Biostatistics, Kaiser Permanente Washington Health Research Institute, Seattle WA, United States

When an important variable in an observational study is hard to measure, an appealing strategy is to predict its values from other variables. This induces Berkson measurement error, the implications of which may not be widely understood. We discuss the case when the predicted variable is an exposure variable in a regression model, with the exposure-outcome association being of interest. Berkson measurement error arises in the regression calibration approach to adjust for measurement error in a regression covariate, for which the unobserved error-free exposure is replaced with a prediction of its expected value given the observed covariates. While this approach can mitigate the bias in the regression parameter induced by the error-prone covariate that is subject to non-differential error, there are several common pitfalls in its application that may cause estimation bias or yield inappropriate standard errors. Additional considerations are needed when a predictor in the calibration equation is also a mediator of the exposure effect in the outcome model.

Mini Symposia

TG4

Comparison of Multivariable Fractional Polynomials with Splines and Penalised Splines in examples

Aris Perperoglou[1], Daniela Dunkler[2], Christine Wallisch[2], Matthias Schmid[3], Patrick Royston[4], Willi Sauerbrei[5]

[1] Newcastle University, United Kingdom, [2] Medical University of Vienna, Austria, [3] Ulm University, Germany, [4] University College London, United Kingdom, [5] University of Freiburg, Germany

When working with multivariable explanatory models, two related challenges are present: select the variables which should be included in the model and what is appropriate functional form. Splines and Fractional Polynomials (FPs) are two flexible tools widely used in practical applications. For FPs Sauerbrei and Royston (1999) described an algorithm that can deal both with variable selection and identification of functional forms in multivariable regression. Multivariable Fractional Polynomials (MFP) combines backward elimination, allowing re-inclusion of omitted variables, with the selection of the functional form for continuous predictors.
For spline users there is no established approach that can be used in a multivariable setting. Hastie and Tibshirani (1990) suggested an approach that decides on the omission of a variable, inclusion as a linear term or smooth term inclusion with a pre-specified degree of freedom. Royston and Sauerbrei (2007) discussed a related approach, in style similar to their MFP algorithm, that can select amongst a choice of pre-specified degrees of freedom.
Alternatively, working with thin-plate regression splines or p-splines degrees of freedom do not have to be pre-defined by the user; typically, a user will select a flexible setting for initial fitting with a penalty term that can control the flexibility of the spline. The smooth term can be defined using generalised cross validation or similar approaches.. For variable selection, Marra and Wood (2011) suggested the inclusion of a second penalty term that can act directly on the null space of the smooth term. When all smoothing parameters tend to zero the second penalty will force the variable completely out of the model.
In this work we present applications of these methods on a set of real data. We consider fractional polynomials, b-splines and natural cubic splines as well as thin-plate regression splines and p-splines to describe the functional form of covariates, combined with aforementioned approaches for variable selection. We discuss practical information of applying the methods, highlight pros and cons and present results.

REFERENCES:
[1] Sauerbrei, Willi, and Patrick Royston. «Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials.» Journal of the Royal Statistical Society: Series A (Statistics in Society) 162.1 (1999): 71-94.
[2] Hastie, Trevor J., and Robert J. Tibshirani. Generalized additive models. Chapman and Hall, 1990. Royston, Patrick, and Willi Sauerbrei. «Multivariable modeling with cubic regression splines: a principled approach.» The Stata Journal 7.1 (2007): 45-70.
[3] Marra, Giampiero, and Simon N. Wood. «Practical variable selection for generalized additive models.» Computational Statistics & Data Analysis 55.7 (2011): 2372-2387.

TG2

Validation of prediction models in the presence of competing risks

Nan van Geloven[1], D Giardiello, EF Bonneville, L Teece, CL Ramspek, M van Smeden, KIE Snell, B van Calster B, M Pohar-Perme, RD Riley, H Putter, E Steyerberg

[1] Leiden University Medical Center, Netherlands

Thorough validation is pivotal for any prediction model before it can be advocated for use in medical practice. For time-to-event outcomes such as breast cancer recurrence, death from other causes is a competing risk. Model performance measures must account for such competing events. We observed that validation guidance for time-to-event outcomes with competing risks is currently spread out over many technical papers, which hampers the uptake of appropriate methods in medical research. We made an accessible overview of contemporary performance measures for time-to-event outcomes with competing risks [1].
In this talk, I present the overview of performance measures for this competing event setting, including the calculation and interpretation of statistical measures for calibration, discrimination, overall prediction error, and clinical usefulness by decision curve analysis. Methods are illustrated for patients with breast cancer, with publicly available data and R code. I will also reflect on the process of writing a tutorial paper on technical estimators for a clinical audience (i.e., STRATOS level 1).

REFERENCES:
[1] van Geloven N, Giardiello D, Bonneville EF, Teece L, Ramspek CL, van Smeden M, Snell KIE, van Calster B, Pohar-Perme M, Riley RD, Putter H, Steyerberg E, on behalf of the STRATOS initiative. Validation of prediction models in the presence of competing risks: a guide through modern methods. BMJ 2022;377:e069249 http://dx.doi.org/10.1136/bmj-2021-069249

TG6 TG8

Use of data-driven simulations to inform real-world survival analyses

Michal Abrahamowicz[1], Marie-Eve Beauchamp[1], Anne-Laure Boulesteix[2], Tim P. Morris[3], Willi Sauerbrei[4], Jay S. Kaufman[1]

[1] McGill University, Montreal, Canada, [2] LMU Munich, Germany, [3] MRC Clinical Trials Unit at UCL, United Kingdom, [4] University of Freiburg

Our overarching goal is to promote the use of targeted simulations to explore particular analytical limitations of a specific real-world prognostic or pidemiological study. Here, we focus on two challenges often encountered in real-world time-to-event (survival) analyses and illustrate how data-driven simulations can yield new insights. The first illustration concerns assessing the impact of omitting an important prognostic factor (cancer stage) on the results of multivariable Cox proportional hazards (PH) analyses used to explore the independent association of the exposure (colon obstruction) with the hazard of all-cause mortality, among persons diagnosed with colon cancer. We show how data-driven simulations permit assessing the joint impact of (i) unmeasured confounding bias and (ii) non-collapsibility, while separating their effects. The second illustration focuses on the pharmaco-epidemiological study of the association between recent use of sedative medications, modeled as a time-varying exposure, and the incidence of cognitive dysfunction. The event is interval-censored as it can be detected only at discrete times of medical visits. Simulation results indicate how the strength of a systematic bias toward the null depends on the frequency of visits and on the assumed strength of the association of interest. For both examples, we designed data-driven simulations to accurately reflect the salient characteristics of the corresponding real-world dataset. We then used the permutational algorithm, validated for simulating event times conditional on time-varying exposures and effects [1], to preserve both the empirical distribution of event times and the assumed or observed hazard ratios for all relevant variables. Our methods and results extend simulation-based quantitative bias analyses to multivariable time-to-event analyses with time-varying exposures.

REFERENCES:
[1] Sylvestre & Abrahamowicz, Statistics in Medicine 2008; 27: 2618-2634.

Why I use Stan

Duco Veen[1]

University Medical Center Utrecht / Utrecht University, Netherlands

In this talk I will first reflect on my own reason to switch to stan, my model was complex in comparison to the amount of data that was available. This is a very common problem and there are many reasons why we can't or don't want to get more data. Thereafter I will reflect on why I kept on using stan and decided to start contributing. I'll provide examples to show how easy it can be to start using stan and provide insight into the ecosystem that comes with stan. Using stan can be beneficial to you in case you have simple model, and especially in case you have complex models. In both cases it is made very ease to check your models, create figures and tables and communicate your results to the outside world. Finally, I'll provide guidance on where to find resources to start and how to join the active community.

Mini Symposia

Why I use Julia

Maren Hackenberg[1]

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

Over the years, many software tools have been developed for supporting statistical modelling and scientific computing in general. While many of these share a set of similar characteristics, such as straightforward manipulation of vectors, matrices and data frames, each approach still carries a legacy of the context in which it has originally been developed.
For example, systems like R and SAS have their origins as convenient interfaces to numerical libraries and have therefore not been optimised for implementing complex algorithms in the language itself, e.g., when constructs such as loops or function calls with large data volumes are needed. On the other side, solutions such as Python come from general-purpose computing and have only been extended post hoc for statistical modelling, in addition to also suffering from similar performance problems. These legacy characteristics become even more challenging with new emerging modelling approaches and tools such as differentiable programming, which, e.g., require straightforward access to automatic differentiation frameworks.
As a rather young language, Julia is not burdened by such legacies and in particular, it has become a key tool for current trends such as differentiable programming. Naturally, the price for this is a still immature package ecosystem.
I will show in several examples how Julia allows to more readily implement scientific ideas that researchers might not have pursued with tools such as R and Python because of runtime inefficiencies or missing flexibility of libraries, e.g., for automatic differentiation. With this, I will argue that Julia removes barriers that so far limit scientific ideas in the modelling community. My examples will focus on differentiable programming, in particular combining deep learning and dynamic modelling, but I will also illustrate in less complex modelling settings how the low-barrier approach of Julia allows for quick prototyping of a prediction model for speedy results under time pressure.
Thus, I hope to convince you to also add Julia to your toolbox for expanding the universe of modelling approaches available to your research.

Why I use Python

Paul Zivich[1]

[1] *University of North Carolina at Chapel Hill, United States*

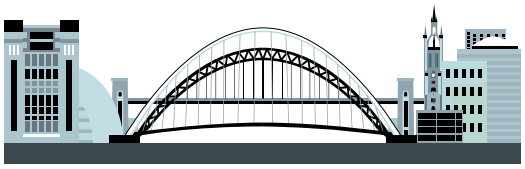
The Python is a high-level, object-oriented, open-source programming language for general-purpose software development. Its free availability, versatility, and large user-base make it an indispensable tool to a variety of data science and machine learning applications. Despite this popularity, Python has been less readily adopted by epidemiologists and biostatisticians. In this talk, I will outline the reasons that Python is my go-to software for both my applied and methodological research projects. Key benefits include readability stemming from the syntax structure, class objects that store both attributes and methods, scalability across systems, ability to interact with other programs, availability of resources for learning, tools to manage multiple versions, and the package ecosystem. To further highlight advantages of Python, I will provide several demonstrations for commonplace statistical applications, including linear regression, inverse probability weighting, and Cox proportional hazards regression. I will then extend to more novel applications, including training a recurrent neural network via abstracts queried from PubMed and generating data with generative adversarial neural networks for simulation studies.

Professional software development: Tips and tricks

Yulia Marchenko[1]

[1] *StataCorp LLC*

In this presentation, I will talk about professional software development and the challenges of producing and supporting a statistical software package. I will share a few tips for how to produce high-quality software, including verification, certification, and reproducibility of the results, and for how to write efficient and stable code. I will also discuss some of the aspects of commercial software development such as clear and comprehensive documentation, consistent specifications, concise and transparent output, extensive error checks, and more.



ISCB43

NEWCASTLE, UK | 2022