MEMTAB 2025

Methods for Evaluating Models, Tests And Biomarkers

Conference Details & Programme

7th International Conference

29th April to 1st May 2025

University of Birmingham, UK

UNIVERSITY OF BIRMINGHAM

Contents

If viewing electronically, please click on headings to move to the appropriate section.

A warm welcome to MEMTAB 2025, at the University of Birmingham!

We are delighted to have you with us as MEMTAB returns to Birmingham, where it began in 2008. This is a critical time for our field and methodology community. With a surge of healthcare studies proposing innovative models, tests and biomarkers, it's crucial to ensure methodological rigour and a positive real-world impact. This year's theme "Methodology That Stands the Test", reflects our commitment to this goal.

Over the next few days, we've got an exciting programme of invited speakers, contributed oral sessions and posters. Participants will disseminate new methods and better research standards, and direct how methodology (and methodologists) should be at the forefront of robust research, implementation and regulation.

Enjoy this enriching time together! Make time to meet, eat, drink and dance (a MEMTAB tradition). Let's challenge and learn from each another in a supportive and constructive environment, united by our shared goal of placing methodology at the heart of our field.

Chair of the Scientific Committee

Gold Sponsors

BIVDA

SMART DATA
ANALYSIS AND STATISTICS

Silver Sponsor

NIHR | National Institute for Health and Care Research

Organising Committee

Richard Riley
Professor of Biostatistics

Kym Snell
Associate Professor in Biostatistics

Jon Deeks
Professor of Biostatistics

Lucinda Archer
Assistant Professor in Biostatistics

Jac Dinnes
Senior Research Fellow

Joie Ensor
Associate Professor in Biostatistics

Beth Hillier
Research Fellow in Biostatistics

Katie Scandrett
Research Fellow in Biostatistics

Hayley Walton
Senior Administrator
(local organising committee member)

Rebecca Whittle
Research Fellow in Biostatistics

Key Information

Pre-conference Course Venue

Murray Learning Centre

University of Birmingham, Birmingham, B15 2FG

Located R28 on the Campus Map

Conference Venue

The Alan Walters Building

University of Birmingham, Birmingham, B15 2TT

Located R29 on the Campus Map

Conference Registration and Query Desk

Day	Opening Times	Venue
Tuesday 29th April	09:30 - 17:00	Murray Learning Centre
Wednesday 30th April	08:30 - 17:30	The Alan Walters Building
Thursday 1st May	08:30 - 17:00	The Alan Walters Building

Conference Social Events

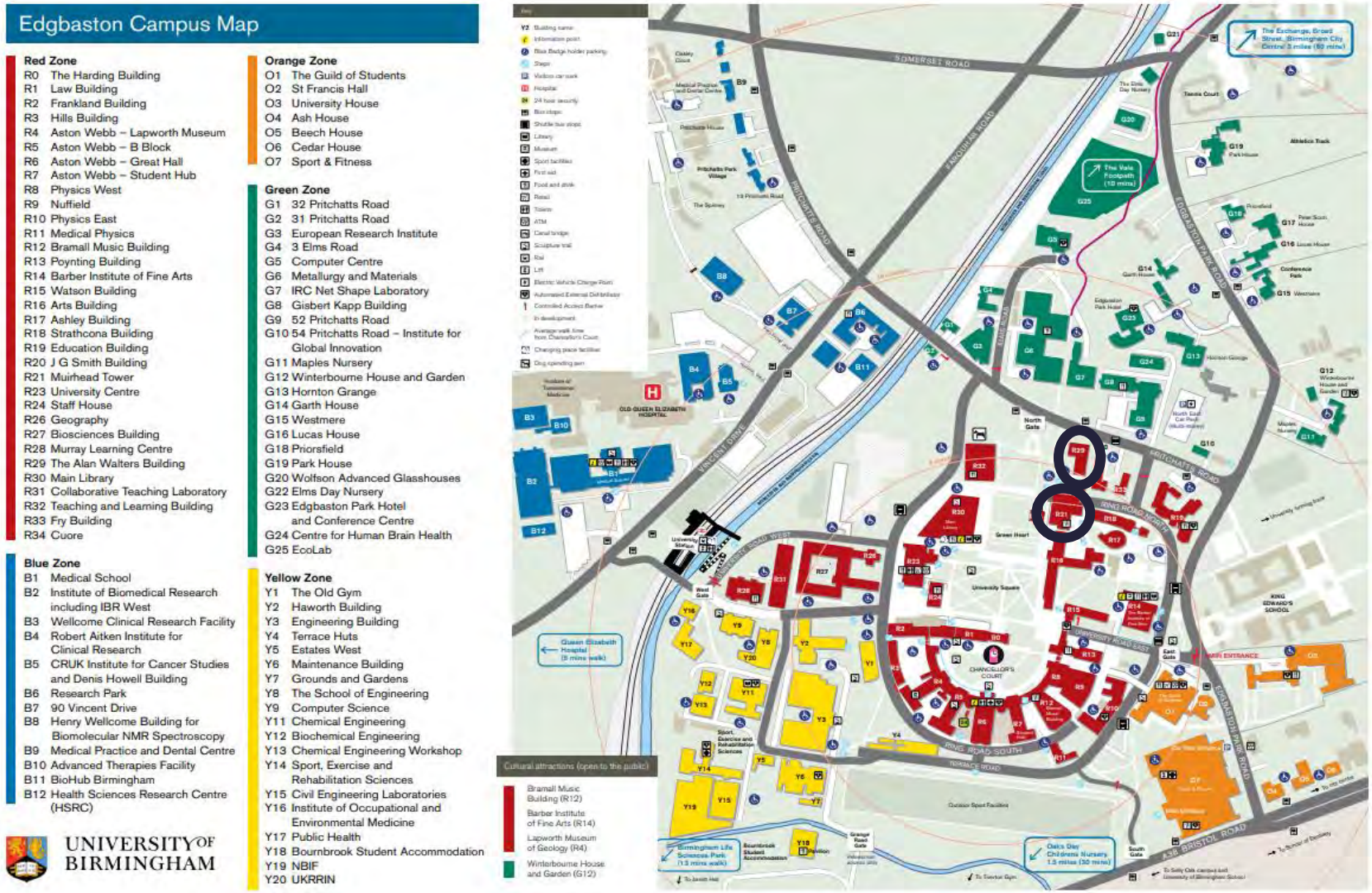
Event	Date/ Time	Venue
Welcome Reception	**Tuesday 29th April** 18:30 – 20:00	Lloyd Suite Edgbaston Park Hotel 53 Edgbaston Park Rd Birmingham B15 2RS
Conference Dinner	**Wednesday 30th April** 19:00 - 23:55	Council House Victoria Square Birmingham B1 1BB

Campus Map

Majority of the MEMTAB 2025 Conference will be taking place in The Alan Walters Building. However, 'Lecture Theatre 2' will be Room G15 in the Muirhead Tower.

There will be staff as well as signage, to guide you from room to room. Luggage storage will be available, if required.

The campus map can also be viewed online



Travel Information

The University of Birmingham's leafy campus is located in Edgbaston, just two miles south-east of the city centre.

The MEMTAB 2025 committee are committed to reducing the environmental impact of the conference wherever possible. Please help us by considering your method of travel carefully.

Rail

The University Train Station is based right on campus and is only 10-minute journey from New Street Station.

The centre of the main campus is a five-minute walk from University station.

Timetables and service updates are available from the National Rail website.

Air

Birmingham Airport has its own convenient train station, Birmingham International, which operates a high-frequency service into Birmingham New Street in under 15 minutes. Change here and follow the directions above to University station.

Taxi

There are taxi ranks at New Street Station, Birmingham Airport and throughout the city centre. The journey to the University takes around 10 minutes, depending on traffic. If you hail a cab, you may need to pay the driver in cash.

For more travel options, please follow this link to the Travel tab from the MEMTAB 2025 website.

What to do on Campus

On campus, the university offers a wide range of attractions for you to enjoy during your stay, all of which boast great history and culture.

Lapworth Museum of Geology

Enabling visitors to explore life over the past 3.5 billion years, the Lapworth Museum showcases exceptional objects from one of the UK's most outstanding geological collections, with state-of-the-art galleries and a range of innovative and interactive exhibits - all completely free of charge. From rocks and fossils to volcanoes, earthquakes, and even dinosaurs, the Museum captures the imagination of all ages.
Located within the Aston Webb A Block building - the building is marked as R4 on the campus map
Opening Times: Monday - Friday - 10:00 - 17:00, Saturday and Sunday - 12:00 - 17:00
Admission Free

Winterbourne House and Garden

Restored to its Edwardian Arts and Craft splendour, Winterbourne House is a unique heritage attraction set within seven acres of beautiful botanic gardens. Winterbourne is a hidden gem, home to beautiful antiques and over 6,000 plant species from around the world. Wander along the woodland walk, stroll through the hazelnut tunnel, cross the 1930's Japanese Bridge or simply soak up the tranquillity of this perfectly English Edwardian home. Located on Edgbaston Park Road, a few minutes' walk from Edgbaston Park Hotel.
Opening Times: 10:30am – 5.30pm,delegates are entitled to a 50% discount on the admission fee.

University of Birmingham Blue Plaque and Sculpture Trails

There have been many influential achievements by brilliant men and women who have worked at the University of Birmingham since its earliest days. The Blue Plaques highlight these special achievements and celebrate those who have helped to shape our heritage as a research university.

The Campus Sculpture Trail allows you to explore the range of styles, subjects and shapes of sculpture on the University's Edgbaston campus. The Faraday Bronze Sculpture was commissioned to mark the centenary of the University of Birmingham's Royal Charter, this is located near the train station.

Want to explore Birmingham further?

Please see our Birmingham & Beyond page on the website.

Conference Programme

Tuesday 29th April - Pre-conference Courses (pre-bookings only)

Murray Learning Centre

Course 1 Room UG06	**An Introduction to Clinical Prediction Models and Sample Size Calculations for Model Development & Evaluation** *Faculty includes Joie Ensor, Kym Snell, Lucinda Archer, Rebecca Whittle, Amardeep Legha and Richard Riley from University of Birmingham*
09:30 - 10:00	Registration and Refreshments
10:00 - 17:00	Course – including lunch

Course 2 Room UG07	**Systematic Reviews of Prognosis Studies** *Faculty includes Anneke Damen and Carl Moons from UMC Utrecht*
09:30 - 10:00	Registration and Refreshments
10:00 - 17:00	Course – including lunch

Course 3 Room UG09	**The Potential and Pitfalls of Predicting Treatment Effects** *Faculty includes David M. Kent from Tufts Medical Center and David Van Klaveren, Erasmus MC University Medical Center*
12:30 - 13:00	Registration and Lunch
13:00 - 17:00	Course

18:30 - 20:00	Welcome Reception Drinks and canapés Lloyd Suite, Edgbaston Park Hotel

Wednesday 30th April - Conference Day 1

The Alan Walters Building

(Registration, Lecture Theatre 1 & 1b, Posters & Refreshments)

and Muirhead Tower (Lecture Theatre 2)

Please note that all plenary sessions in Lecture Theatre 1 will also be streamed into adjacent room, Lecture Theatre 1b (floor 1).

08:30 - 09:30	**Registration and Refreshments**	
09:30 - 09:40	Welcome and Introduction - *Lecture Theatre 1 (The Alan Walters Building)*	
09:40 - 10:25	**Session 1: S1. Methodology that stands the test** *Patrick M Bossuyt, Professor of Clinical Epidemiology, University of Amsterdam* Chaired by Richard Riley - *Lecture Theatre 1 (The Alan Walters Building)*	
10:25 - 11:00	Break	
	Session 2: Methodology for models, tests and biomarkers	
	Tests Chaired by Jac Dinnes *Lecture Theatre 1* *(The Alan Walters Building)*	**Predictions** Chaired by Lucinda Archer *Lecture Theatre 2* *(Muirhead Tower)*
11:00 - 11:15	**O1. QUADAS-3: updated tool to evaluate risk of bias and applicability concerns in diagnostic test accuracy studies** *Penny Whiting, University of Bristol*	**O6. Performance evaluation of predictive AI models to support medical decisions: overview and guidance** *Ben Van Calster, KU Leuven*
11:15 - 11:30	**O2. What is the evidence base for claims of accuracy for rapid self-test diagnostics sold in UK retail settings?** *Beth Hillier, University of Birmingham*	**O7. A software implementation for sample size calculation targeting precise risk predictions** *Joie Ensor, University of Birmingham*

11:30 - 11:45	**O3. Research Waste in Evidence Synthesis for Health Population Screening: A Systematic Review** *Sarah Batson, University of Warwick*	**O8. Developing a clinical prediction model with a continuous outcome: sample size calculations to target precise predictions** *Rebecca Whittle, University of Birmingham*
11:45 - 12:00	**O4. Bayesian statistical methods for diagnostic studies that allow early termination for futility** *Jordan Oakley, Newcastle University*	**O9. Adapting sample size calculations for the development of prediction models to control for model stability** *Menelaos Pavlou, University College London*
12:00 - 12:15	**O5. Opportunities to speed up in-vitro diagnostic adoption and patient access in the UK: the pre-eclampsia testing timeline** *Katie Scandrett, University of Birmingham*	**O10. How to Handle Missing Data across the Development, Validation and Implementation of Clinical Prediction Models** *Glen Martin, University of Manchester*
12:15 - 13:15	Lunch and Networking	
13:15 - 13:55	**Session 3: S2. Value of information analysis: towards a value-based approach in biomarker and prediction model research** *Mohsen Sadatsafavi, Associate Professor, University of British Columbia* Chaired by Joie Ensor - *Lecture Theatre 1 (The Alan Walters Building)*	
13:55 - 14:05	Break	
	Session 4: Methodology for models, tests and biomarkers	
	Tests Chaired by Hans Reitsma *Lecture Theatre 1* *(The Alan Walters Building)*	**Prediction** Chaired by Carl Moons *Lecture Theatre 2* *(Muirhead Tower)*

14:05 - 14:20	**O11. Simulation Study Examining Impact of Study Design Factors on Variability Measures** *Laura Quinn, University of Birmingham*	**O16. Value-of-Information Analysis for External Validation of Risk Prediction Models in Multicenter Studies and Systematic Reviews** *Laure Wynants, Maastricht University*
14:20 - 14:35	**O12. Real world implementation of the Biomarker Toolkit: a Tool aiming to quantifiably assess biomarker utility and guide development** *Katerina-Vanessa Savva, Imperial College London*	**O17. Comparing Performance of Methods that Correct for Data Distribution Shift when Developing Clinical Prediction Models: A Simulation Study** *Haya Elayan, University of Manchester*
14:35 - 14:50	**O13. Methodology to create evidence-based testing panels for monitoring long-term conditions in primary care** *Martha Elwenspoek, University of Bristol*	**O18. Use of statistical process control to monitor calibration-in-the-large of a clinical prediction model** *David Jenkins, University of Manchester*
14:50 - 15:05	**O14. Measurement Error: Unlocking Estimates of Test Variability From Routine Data. Methods for Statistical Analysis and a Case-Study Series** *Simon Baldwin, University of Birmingham*	**O19. Combining calibration plots from multiple centers or datasets** *Lasai Barreñada, KU Leuven*
15:05 - 15:20	**O15. Evaluation of diagnostic tests with spatially or temporally clustered data, part 1: The choice of estimands and estimators affects results and interpretation** *Nicole Rübsamen, University of Münster*	**O20. Network meta-analysis of prediction models using aggregate or individual participant data - A scoping review and recommendations for reporting and conduct** *Maerziya Yusufujiang, UMC Utrecht*

15:20 - 16:15	Break and Poster Viewings – *Room G11 (The Alan Walters Building)*
	Session 5: Patient and Public Involvement & Engagement Chaired by Kym Snell - *Lecture Theatre 1 (The Alan Walters Building)*
16:15 - 16:30	**S3. The need for PPIE within methodology research** *Laura Gray, Biostatistics Research Group, University of Leicester*
16:30 - 16:45	**S4. Experiences of a PPIE representative within methodology research** *Emily Lam, PPIE Representative*
16:45 - 17:00	**S5. Establishing and working with PPIE panels from prediction model research: what we have learnt** *Paula Dhiman, Senior Research in Medical Statistics, University of Oxford*
17:00 - 17:15	**S6. 'PPIE meets statistics': educating PPIE groups about prediction models and research methodology** *Pradeep Virdee, Nuffield Department of Primary Care Health Sciences, University of Oxford*
17:15 - 17:30	**Open Discussion** - *Lecture Theatre 1 (The Alan Walters Building)*
17:30	Close - *Lecture Theatre 1 (The Alan Walters Building)*

19:00 - 23:55	**Conference Dinner** (prebooked tickets only) Council House, Birmingham

Thursday 1st May - Conference Day 2

The Alan Walters Building (Lecture Theatre 1 & 1b, Posters & Refreshments) and Muirhead Tower (Lecture Theatre 2)

08:30 - 09:00	**Registration and Refreshments**	
09:00 - 09:05	Welcome to Day 2 - *Lecture Theatre 1 (The Alan Walters Building)*	
09:05 - 09:50	**Session 6: S7. Are AI-enabled systems in healthcare fit for purpose? Toward equitable, fair and trustworthy systems for disease detection and risk prediction** *Alicja Rudnicka, Professor of Statistical Epidemiology in the Population Health Research Institute, City St Georges, University of London* Chaired by Joie Ensor - *Lecture Theatre 1 (The Alan Walters Building)*	
09:50 - 10:45	Break and Poster Viewings – *Room G11 (The Alan Walters Building)*	
	Session 7: Methodology for models, tests and biomarkers	
	Tests Chaired by Rafael Perera *Lecture Theatre 1 (The Alan Walters Building)*	**Prediction** Chaired by Paula Dhiman *Lecture Theatre 2 (Muirhead Tower)*
10:45 - 11:00	**O21. Evaluating Diagnostic Tests Against Composite Reference Standards: Quantifying and Adjusting for Bias** *Vera Hudak, University of Bristol*	**O27. PROBAST+AI: An updated quality, risk of bias and applicability assessment tool for prediction models using regression or artificial intelligence methods** *Anneke Damen, UMC Utrecht*

11:00 - 11:15	**O22. Improving the reference standard in diagnostic accuracy studies: Evaluating a latent class model against a panel of expert clinicians** *Tom Parry, University College London*	**O28. Guidance for unbiased predictive information for healthcare decision-making and equity (GUIDE): considerations when race may be a prognostic factor** *David Kent, Tufts Medical Center*
11:15 - 11:30	**O23. Examining the Association between Estimated Prevalence and Diagnostic Test Accuracy Using Directed Acyclic Graphs** *Yang Lu, McGill University*	**O29. A simulation study investigating the impact of the prediction paradox on clinical prediction model performance** *Samantha Pacynko, University of Manchester*
11:30 - 11:45	**O24. Diagnostic accuracy of tests for SARS-CoV-2 acute infection: Distinguishing measurands from target conditions** *Joanna Merckx, McGill University*	**O30. CHARIOT: A prediction-under-intervention model for cardiovascular primary prevention** *Matthew Sperrin, University of Manchester*
11:45 - 12:00	**O25. The estimand framework for diagnostic accuracy studies** *Antonia Zapf, University Medical Center Hamburg-eppendorf*	**O31. Stronger penalties on treatment-covariate interactions improve treatment effect predictions and prevent potential treatment mistargeting** *David Van Klaveren, Erasmus MC*
12:00 - 12:15	**O26. How do authors of comparative accuracy studies analyse data when reporting a comparative conclusion: methodological review?** *Yaxin Chen, Amsterdam UMC*	**O32. Effects of Using Natural Language Processing for Cohort Selection from Electronic Health Records on Subsequent Prognostic Prediction Model Performance** *Isa Spiero, UMC Utrecht*
12:15 - 13:15	Lunch and Networking	

13:15 - 13:50	**Session 8: S8. Performance Evaluation of Diagnostics - Industry Challenges and Opportunities for Regulatory Science** *Mike Messenger, BIVDA* Chaired by Clare Davenport - *Lecture Theatre 1 (The Alan Walters Building)*
13:50 - 14:00	Break
	Session 9: Regulation of Tests and Models Chaired by Niels Peek - *Lecture Theatre 1 (The Alan Walters Building)*
14:00 - 14:15	**O33. Identifying Priority Areas for Target Product Profile Development in Early Cancer Diagnostics** *Bethany Shinkins, University of Warwick/ NICE*
14:15 - 14:30	**O34. Developing diagnostic target product profiles for managing infections and exacerbations in cystic fibrosis: a sequential mixed-methods design.** *Nicola Howe, Newcastle University*
14:30 - 14:45	**O35. Lost in Translation: The Current and Future Regulatory Landscape as an Often-Overlooked Hurdle for Impact in Clinical Prediction Models** *Benjamin Perry, University of Birmingham*
14:45 - 15:00	**O36. Assessment of Prediction Models in Europe: Gaps in Evidence Requirements** *Tuba Saygin Avsar, NICE*
15:00 - 15:40	Break
15:40 - 16:40	**Session 10: Standing The Test For The Future** Panel: Jon Deeks, Rishi Gupta, Anne de Hond, Chris Hyde, Mariska Leeflang, Sowmiya Moorthie Chaired by Richard Riley - *Lecture Theatre 1 (The Alan Walters Building)*
16:40 - 17:00	Awards and Closing Remarks - *Lecture Theatre 1 (The Alan Walters Building)*

Invited Speakers

Prof. Patrick M Bossuyt

Prof. Patrick M Bossuyt is the professor of Clinical Epidemiology at the Amsterdam University Medical Centers, where he leads the Biomarker and Test Evaluation research program.

The BiTE program aims to appraise and develop methods for evaluating medical tests and biomarkers, with an emphasis on clinical performance, and to apply these methods in relevant clinical studies. In doing so, the program wants to strengthen the evidence base for rational decision-making about the use of tests and testing strategies in health care. Prof. Bossuyt spearheaded the STARD initiative for the improved reporting of diagnostic test accuracy studies. Prof. Bossuyt has authored and co-authored several hundred publications in peer reviewed journals and serves on the editorial board of a number of these, including Radiology and Clinical Chemistry. He acted as chair of the Department of Clinical Epidemiology & Biostatistics at his university, chaired the Division of Public Health, and was Dean of Graduate Studies. For 10 years, Prof. Bossuyt also chaired the Scientific Advisory Committee of the Dutch Health Insurance Board, which oversees the health care benefits covered in the national health insurance program. In 2024, he received the ADLM Wallace H. Coulter Lectureship Award. This award "recognizes an outstanding individual who has demonstrated a lifetime commitment to, and made important contributions that have had a significant impact on education, practice and/or research in laboratory medicine or patient care."

Dr. Paula Dhiman

Paula is a Senior Research Fellow in Medical Statistics based in the Centre for Statistics in Medicine at the University of Oxford.

Her research also includes reviewing the quality of medical research studies, informing the development of reporting guidelines and developing methodological guidance, all with a focus on prediction modelling. She is currently funded by CRUK to investigate sample size requirements when using machine learning in prediction model research.

Paula is a member of the UK EQUATOR Network (promoting the use of reporting guidelines for Enhancing the Quality and Transparency Of health Research) and the TRIPOD group which produced the reporting guideline for prediction modelling studies (TRIPOD+AI). Paula is the current appointed Chair of the QResearch Scientific Committee which reviews applications to one of the UKs largest primary care databases and the statistical lead for the Blood and Transplant Research Unit for Data Driven Transfusion Practice. She is also the Nigel James Junior Research Fellow in Medical Statistics at Pembroke College, Oxford.

Professor Laura Gray

Professor Laura Gray co-leads the Biostatistics Research Group at University of Leicester.

Laura has a strong and sustained track-record of leading cross-disciplinary research spanning methodological and clinical areas, with a focus on identification, prevention and management of chronic and multiple long-term conditions. Her methodological research is motivated by practical problems encountered in her applied health research; specialising in trials, prognostic modelling and evidence synthesis. Laura has a passion for ensuring meaningful patient and public involvement in her research, and has led a number of research studies in this area.

Emily Lam

I am a retired Chinese female living with several chronic health conditions in a rural part of Cheshire. Before my retirement, I had a long career in nursing across primary care and various specialities in hospitals. I have an Open University degree in research methodology and also worked as a senior nurse in research and development for a few years. After retirement, I set up and ran an education recruitment business for fourteen years. Then I participated as an Expert By Experience in many Care Quality Commission inspections until the pandemic struck. The work took me to hospitals, care homes and GP surgeries across the region to interview patients and carers, and led me to take a keen interest in the value of people's insights and experiences, especially those from deprived communities. I developed a passion to advocate for them, for instance, in my capacity as a lay member on NICE committees which I served for over ten years. In recent years, I have taken an interest in research methodologies and thought about the public and patient's role in their work. I believe better research will facilitate better outputs to shape health care policies. We need to bring all the skills, insights and understanding of people from different cultures in our society which already has a deepening divide between the wealthy and the poor threatening health inequalities. I hope that by bringing our different perspectives and sharing our understanding with one another, we can be working together to better shape healthcare for public good.

Professor Michael Messenger

Head of Regulatory Strategy at BIVDA, Director at Insightful Health Ltd and Visiting Professor at the University of Leeds

Mike has >20 years experience in the medical technology sector, focused primarily on Diagnostics. He holds a visiting chair at the University of Leeds and is an expert in diagnostics regulation, evaluation and market access strategy. He is an advisor to a diverse international clientele, ranging from government regulators and global health organizations, through to donors and industry. Recent highlights include drafting WHO's Implementation Plan for Pandemic Influenza Preparedness and developing a portfolio of Target Product Profiles and Technical Specifications for FIND, GAVI and WHO.

Previously Head of Diagnostics at the MHRA and a Senior Scientific Advisor to DHSC during the pandemic, he led COVID IVD EUAs, development of Target Product Profiles, NICE assessments, advised government test validation groups and the 100-day mission, co-developed the NIHR CONDOR trial, and advised national and EU agencies.

Mike previously established the Leeds Centre for Personalised Medicine and Health and the NIHR In-Vitro Diagnostic Co-Operatives in Leeds. He served for 10 years on the NICE Diagnostic Advisory Committee and is currently a member of International Standards Organisation (ISO) TC 212 "Medical laboratories and in vitro diagnostic systems" working groups 1, 3 & 4 and British Standards Institute (BSI) CH 212.

Professor Alicja Rudnicka

Professor Alicja Rudnicka is a Professor of Statistical Epidemiology in the Population Health Research Institute, City St Georges, University of London, and co-founder member of the Artificial Intelligence and Automated Retinal Image Analysis Systems ARIAS Research Group.

Recent work has focussed on the use of AI methods for analysing retinal images in relation to (i) creation of a retinal microvascular phenotypes for disease risk prediction and (ii) performance equity of AI for detection diabetic eye disease from retinal images, for potential deployment within the English NHS National Diabetic Eye Screening Programme. Public, patient and practitioner engagement and involvement has also formed a growing area of research. She works as part of a multi-disciplinary national and international team, which includes AI/computer scientists, clinicians, data scientist, behavioural scientists and epidemiologists from different academic institutions.

Dr. Mohsen Sadatsafavi

Associate Professor and Associate Director of Research at the University of British Columbia's Faculty of Pharmaceutical Sciences.

Mohsen is an academic epidemiologist with a keen interest in the methods and applications of decision theory in healthcare and health research. He is particularly interested in improving efficiency in medical decision-making and evidence collection. He studies the implications of uncertainty in evidence and its impact on the outcomes of decisions.

Mohsen believes that questions around decision efficiency are fundamentally similar at clinical and health policy levels but observes a lack of interaction between the respective methodological communities. Through his methodological work on uncertainty quantification and Value of Information analysis, he aims to bridge this gap. His applied research focuses on respiratory diseases. He has authored / co-authored over 270 publications, serves on the Scientific Committees of several international studies, and has contributed to guideline development in respiratory medicine. Mohsen is currently the Statistical Editor for Thorax and an Editorial Board member of Medical Decision Making.

Dr Pradeep Virdee

Pradeep is a Senior Medical Statistician based in the Cancer Theme at the Nuffield Department of Primary Care Health Sciences, University of Oxford.

Pradeep's expertise include the acquisition and use of linked electronic health records for large-scale cancer diagnostic studies and prediction modelling studies, utilising repeated measures data for clinical risk prediction, and leading statistical components of many prospective clinical trials. His interests include improving efficiency in statistical programming and management of big data and methodological innovations for earlier cancer detection. Pradeep proactively engages with patients and members of the public to raise their awareness and understanding of the methodological components of applied research and the interpretation and communication of research.

Invited Speaker Abstracts

In order of appearance in the programme

S1. Methodology that stands the test

Patrick M Bossuyt[1]
[1]Amsterdam University Medical Center

It is now universally acknowledged: medical tests should be carefully evaluated before their use can be recommended in health care. Yet while trials became the cornerstone in the evaluation of pharmaceuticals and other interventions, the methods for evaluating laboratory tests, imaging, prediction models, and other forms of medical testing were trailing in the shadows of drug development. Why the delay? A combination of factors can be held responsible. Regulations that embraced drug oversight were introduced decades earlier, compared to tests. Tests do not directly save lives: there is an indirect link between testing and health outcomes. This presentation will time-travel through the evolution of test evaluation, then zoom forward with a cautious forecast of what is next.

S2. Value of information analysis: towards a value-based approach in biomarker and prediction model research

Mohsen Sadatsafavi[1]
[1]University of British Columbia

Models, tests, and biomarkers are developed and assessed in finite samples. As such, the information they carry is accompanied by uncertainty. The contemporary approach towards uncertainty assessment in predictive analytics is precision-driven, based on reporting confidence intervals around metrics of model performance such as AUC or calibration slope. Accordingly, the design of development, validation, and implementation studies is informed by targeting precision around such metrics.

Ultimately, however, we use diagnostic and prognostic information for patient care.

When faced with the decision whether to use or not use a test, the relevance of precision metrics is doubtful. Decision theory provides an alternative perspective to this problem: Uncertainty is associated with a loss of value because it might prevent us from making the most optimal decision. The impact of uncertainty can thus be expressed in terms of its associated loss in value. This approach towards uncertainty assessment, referred to as Value of Information (VoI) analysis, has been practiced in risk analysis and health technology assessment, but has only recently been applied to predictive analytics.

In this talk, we first review two fundamental VoI metrics: EVPI (Expected Value of Perfect Information) and EVSI (Expected Value of Sample Information) and recent advancements in defining and computing these metrics for development and validation studies. We then offer two actionable suggestions: 1) formal incorporation of EVPI into Decision Curve Analysis in lieu of reporting confidence intervals around net benefit. 2) formal incorporation of EVSI into current power and sample size calculations for development and validation studies. Finally, we discuss multiple open questions in this nascent area, including the incorporation of other sources of uncertainty, VoI analysis for different outcome types and study designs, and the potential of this approach for investigating algorithmic fairness.

We conclude that when clinical utility of models, tests, and biomarkers is concerned, a 'value-based' perspective can complement the contemporary 'precision-driven' approaches towards uncertainty quantification and design of empirical studies.

S3. The need for PPIE within methodology research

Laura Gray[1]
[1]University of Leicester

Patient and public involvement and engagement (PPIE) is well-embedded in applied health care research in the UK, with PPIE a funding requirement. Unlike applied health research, there is little research on how to conduct meaningful PPIE in statistical methodology research. Statistical methodology research involves the development, evaluation or comparison of statistical methods for the design or analysis of research studies. The technical nature of this research and often confusing terminology can make PPIE challenging. The PPI-SMART group at the University of Leicester has developed a number of resources to aid those undertaking PPIE for statistical methodology research based on the needs of statisticians identified in a nationwide survey. In this talk, I will introduce our work to date and demonstrate the positive benefits PPIE can have to statistical methodology research.

S4. Experiences of a PPIE representative within methodology research

Emily Lam

In my talk, I hope to share my perspective as a public contributor interested in supporting methodology research. Coming from a background as a lay member on NICE's Medical Technology Advisory Committee of over ten years, my initial involvement in health has obviously been in clinical research. However, my reflection over the years is that most clinical research started life as an idea or an observation in a laboratory, and robust methodology research is fundamental to ensuring that the designs, tools and analyses developed and produced are effective, reliable and reproducible for knowledge generation. This solid foundation is essential for supporting clinical researchers in winning public confidence. What I am going to talk about can be summed up as asking challenging questions around why we need to involve and engage the public if we aspire to achieve the best outcome in methodology research.

S5. Establishing and working with PPIE panels from prediction model research: what we have learnt

Paula Dhiman[1]
[1]*University of Oxford*

It is strongly encouraged to include patient and public involvement and engagement (PPIE) when developing a research study and prediction model research is no exception. However, where and how PPIE might inform an applied medical prediction model research study is often more obvious than for a prediction methodology study. Further, PPIE groups for prediction model methodology research are also few and far in between.

In this talk, I will share experiences of establishing a prediction model methodology PPIE panel in collaboration with the University of Birmingham. I will include what has been learnt along the way and I will also draw experiences on how I have embedded PPIE for my prediction methodology research.

S6. PPIE meets statistics: educating PPIE groups about prediction models and research methodology

Pradeep Virdee[1]
[1]*University of Oxford*

Collaboration between PPIE contributors and statisticians has historically been sparse. Our recent work identified the lack of familiarity of data and statistics among PPIE contributors to be one key barrier to involvement. The BETTA project at the University of Oxford aims to improve collaboration and includes a training theme that offers in-person events and webinars to train PPIE contributors about data and statistics. I will provide feedback about our first event, including why the project was set-up and established, how I approached the education of statistics/methodology, primarily prediction modelling, and lessons learned and next steps.

S7. Are AI-enabled systems in healthcare fit for purpose? Toward equitable, fair and trustworthy systems for disease detection and risk prediction

Alicja Rudnicka[1]
[1]*University of London*

The presentation will explore the potential of artificial intelligence (AI)-based algorithms and models to transform healthcare, particularly in the areas of disease detection, prediction, and management. Existing guidelines, such as those provided by TRIPOD+AI, provide clear recommendations for the development and evaluation of multivariable prediction models. However, the evaluation of AI systems as medical devices in real-world settings introduces new challenges. This presentation will highlight key study design considerations necessary for conducting trustworthy, equitable, and fair real-world evaluations of AI models/systems within the healthcare context.

S8. Performance Evaluation of Diagnostics - Industry Challenges and Opportunities for Regulatory Science

Michael Messenger[1]
[1]*BIVDA*

Abstract:
Global trends to enhance regulatory oversight of medical devices and IVDs are driving increased requirements for manufacturers to demonstrate the scientific, analytical, and clinical performance of their products. While regulations, such as the EU IVDR and UK MDR2002, outline the key legal requirements, significant methodological uncertainties and challenges persist regarding the generation and evaluation of the necessary evidence in accordance with the current state of the art.

This talk will introduce the requirements for Performance Evaluation of IVDs and highlight current challenges and opportunities for methodological innovation and standardisation.

Oral Presentation Abstracts

O1. QUADAS-3: updated tool to evaluate risk of bias and applicability concerns in diagnostic test accuracy studies

Penny Whiting[1], Miss Eve Tomlinson[1], Dr Bada Yang[2], Dr Clare Davenport[3], Professor Mariska Leeflang[4], Professor Sue Mallett[5], Dr Anne Rutjes[6]

[1]Population Health Science, Bristol Medical School, University of Bristol, [2]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, , [3]Test and Prediction Group, Department of Applied Health Science, University of Birmingham, [4]Amsterdam University Medical Centers, University of Amsterdam, , [5]Centre for Medical Imaging, University College London, [6]UniCamillus - Saint Camillus International University of Health Science, , [7]Dipartimento di Scienze Mediche e Chirurgiche, Materno-Infantili e dell'Adulto, Università di Modena-Reggio Emilia

Background: The QUADAS-2 tool, published in 2011, was designed to evaluate the risk of bias and applicability of diagnostic test accuracy (DTA) studies. The publication reporting QUADAS-2 has been cited over 12, 000 times and it is the recommended tool to assess risk of bias and applicability of studies for major HTA organisations. Although feedback on QUADAS-2 has generally been positive, some signaling questions have been identified as problematic and the tool could be improved based on features included in more recently developed tools.

Objectives: To update QUADAS-2 to develop the new QUADAS-3 tool.

Methods: We established a core-group of methodological experts to lead the development of QUADAS-3 supported by a wider steering group.

We followed the following steps:

- Summarised modifications made to QUADAS-2 for the Cochrane Handbook
- Web-based survey of reviewers that have used QUADAS-2
- Considered developments from more recent tools
- Review of methodological studies that had evaluated QUADAS-2
- Review of 50 Cochrane DTA reviews to highlight challenges with the assessment of applicability

We have produced a draft tool which is currently undergoing piloting. The results of the piloting, which will also include a comparison of the use of signalling questions with signalling statements, will be used to inform the final version of the tool.

Results: The new tool follows a similar structure to the QUADAS-2 tool but with some major updates. Key changes include:

- An option to define separate synthesis questions rather than just a single review question
- A new section on defining the ideal test accuracy trial for each synthesis question
- Assessment of risk of bias and applicability at the accuracy estimate level rather than the study level
- A change in answers to signaling questions to include options of “probably yes” and “probably no” and to replace “unclear” with “no information”
- Replacement of “Flow and Timing domain” with new “Analysis” domain
- Changes to some signaling questions
- Inclusion of a section for judging overall risk of bias and applicability (across domains)

Conclusions: QUADAS-3 will be introduced at the conference and the results of piloting discussed.

O2. What is the evidence base for claims of accuracy for rapid self-test diagnostics sold in UK retail settings?

Beth Hillier[1,2], Simon Baldwin[1,2], Katie Scandrett[1], Ridhi Agarwal[1], Aditya Kale[3], Joseph Alderman[2,3], Trystan Macdonald[2,4], Alex Richter[2,5], Clare Davenport[1,2], Jon Deeks[1,2]
[1]Department of Applied Health Sciences, University of Birmingham, [2]NIHR Birmingham Biomedical Research Centre, University Hospitals Birmingham NHS Foundation Trust and University of Birmingham, [3]Institute of Inflammation and Ageing, University of Birmingham, [4]University Hospitals Birmingham, NHS Foundation Trust, [5]Institute of Immunology and Immunotherapy, University of Birmingham

Background: The availability and public interest in self-tests on the UK high street has grown. Self-tests are regulated in the UK and EU by Notified Bodies who assess clinical and layperson study reports (CSRs and LSRs) provided by manufacturers. There are concerns about the regulatory process and their assessments.

Objectives**:** To review the evidence base and performance claims of self-tests.

Methods**:** Self-tests were obtained in supermarkets, pharmacies and health shops within 10 miles of the University of Birmingham in 2023. Instructions for Use (IFU) and packaging were reviewed. We requested LSRs and CSRs from manufacturers and distributors by email, with follow-up emails for non-repliers. We critically appraised LSRs and CSRs using QUADAS-2.

Results**:** Thirty-five self-tests were identified; thirty were obtained, which used seven types of samples and detected twenty different biomarkers.

IFUs made accuracy claims for 24/30 tests: 19, 17 and 16 for accuracy, sensitivity and specificity, respectively. Performance claims of ≥98% were made on accuracy for 53% (10/19) of tests, 41% (7/17) on sensitivity and 63% (10/16) on specificity. No statements were made on predictive values. Where reported, 28% (5/18) used inappropriate reference standards, comparing self-tests against similar rapid tests.

We obtained CSRs and LSRs for 40% (12/30) of the tests (9 unique documents). Manufacturers refused requests for 7 tests and never replied for 11. Most CSRs were poorly detailed, with participant descriptions not reported in 78% (7/9). Few demographics were presented within LSRs. Details available showed some tests were evaluated using unrepresentative sample groups. LSRs provided limited evidence of real-world usability, with no mention of blinding in 89% (8/9) of studies. Most studies assessed (73%) were rated unclear/high RoB across all QUADAS-2 domains, and most CSRs (78%) had unclear/high applicability concerns across all domains.

Conclusions**:** Much of the evidence behind performance claims for self-tests is not publicly available or of poor methodological quality. Reports lack important information, especially regarding study populations. CSRs are often not conducted in real-world settings, but are laboratory-based and use specimens without information upon whom they are based. The lack of transparency and poor reporting negatively affect the safety, credibility and reliability of self-tests.

References:

[1] BSI: An In Vitro Diagnostics Notified Body. A BSI Guide to the In Vitro Diagnostic Directive. 2012 [Available from: https://www.bsigroup.com/globalassets/localfiles/en-hk/medical%20device/bsi-md-ivd-diagnostic-directive-guide-brochure-uk-en.pdf, accessed Oct 2024].

[2] Gram EG, Copp T, Ransohoff DF, et al. Direct-to-consumer tests: emerging trends are cause for concern. *BMJ* 2024;387:e080460. doi: 10.1136/bmj-2024-080460

[3] Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. *Annals of Internal Medicine* 2011;155(8):529-36. doi: 10.7326/0003-4819-155-8-201110180-00009

O3. Research Waste in Evidence Synthesis for Health Population Screening: A Systematic Review

Sarah Batson[1], Matthew Randell[1], Catherine Bane[1], Julia Geppert[1], Pranshu Mundala[1], Chris Stinton[1], Sian Taylor-Phillips[1]
[1]*University Of Warwick*

Background: Evidence synthesis products, such as systematic reviews, evidence summaries, and evidence maps, play a critical role in informing decisions for population health screening programs and policies worldwide. These products can address a range of research questions, including the incidence or prevalence of conditions, screening test accuracy, benefits and harms of screening and treatment, and the cost-effectiveness of screening. However, many national and international organisations independently assess similar screening programs, raising concerns about unnecessary duplication of evidence synthesis efforts contributing to research waste—defined as inefficient duplication and misallocation of resources. The issue of research waste in evidence synthesis for population health screening has yet to be systematically assessed or empirically evaluated.

Objectives: To estimate the extent of research waste in the production of evidence synthesis products for population health screening, providing an initial assessment of its magnitude and implications.

Methods: Evidence synthesis products supporting screening recommendations for adult populations, published by the UK National Screening Committee (UK NSC) and the US Preventive Services Task Force (USPSTF) between 2014 and 2024, were identified as index reviews. For each index review, Embase, Medline and national and international organisation websites were searched for other reviews on the same topic, defined as addressing the same key research questions with at least partial overlap in the population, interventions, comparisons, and outcomes.

Results: A total of 48 evidence synthesis products covering 33 conditions comprised the index reviews. Overlapping reviews were identified for 85% (41/48) of the index reviews, with a median of 4 additionally identified reviews per index review (range:1-45; IQR: 2-12).

Conclusions: The results of this study highlight research waste due to significant duplication in evidence synthesis efforts. There is an urgent need for national policymakers to collaborate in establishing efficient collaboration practices for evidence synthesis, enabling reuse and adaptation. Such collaboration could include sharing ongoing reviews, conducting multi-region comprehensive reviews, and utilising stratified analyses to tailor data to individual country needs. By adopting these strategies, organisations can streamline the process, reduce unnecessary duplication, improve global policy making efficiency, and redirect resources toward new research priorities and other public health challenges.

O4. Bayesian statistical methods for diagnostic studies that allow early termination for futility

Jordan Oakley[1], Rachel Binks[1,2], Timothy Hicks[2,3], Alison Bray[4], Kile Green[2], Will Jones[3,5], James Wason[6], Kevin Wilson[1]

[1]School of Mathematics, Statistics & Physics, Newcastle University, Newcastle Upon Tyne, United Kingdom, [2]NIHR HealthTech Research Centre (HRC) for Diagnostic and Technology Evaluation (DTE), Translational and Clinical Research Institute, Newcastle University, Newcastle Upon Tyne, United Kingdom, [3]NIHR HealthTech Research Centre (HRC) for Diagnostic and Technology Evaluation (DTE), Newcastle Upon Tyne Hospitals NHS Foundation Trust, Newcastle Upon Tyne, United Kingdom, [4]Northern Medical Physics and Clinical Engineering, The Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, United Kingdom, [5]Faculty of Science and Engineering, Center of Excellence for Data Science, Artificial Intelligence and Modelling, The University of Hull, Hull, United Kingdom, [6]Population Health Sciences Institute, Newcastle University, Newcastle Upon Tyne, United Kingdom

Background: Seamless designs and group sequential designs can be used to make diagnostic and clinical studies more efficient. Group sequential designs allow a study to stop early for futility (or efficacy) at a preplanned interim analysis. Stopping early for futility can prevent exposing patients to an inferior diagnostic procedure or treatment, and make room in the development pipeline for better tests and treatments. Seamless designs aim to re-use data collected in one phase of a study in a later phase of a study. In this scenario, the second phase of a study will only commence if the outcome of the first is satisfactory.

Both designs allow for early termination of the study based on futility criteria. The statistical models used to describe such designs should take account of the data collection procedure so that the model is aligned to the data generating process and can appropriately estimate the parameters of interest.

Objectives: This talk will introduce diagnostic designs with futility criteria and present statistical methods which incorporate this design feature. A Bayesian statistical framework to estimate diagnostic parameters (such as sensitivity or specificity) for designs with posterior-based truncated criteria will be proposed.

Methods: We present a Bayesian procedure to obtain the parameter posterior distribution for studies which allow for early termination using posterior-based truncated criteria. We implement this procedure with commensurate and power priors that allow for the commensurability of the information in the first phase of a study and second phase to determine how much of the first phase's data is used.

Results: Using a simulation study we show the posterior distributions are well-calibrated and we highlight the consequences of ignoring the futility criteria when making inferences. We illustrate the methods using real data used to develop a biomarker test for ventilator associated pneumonia.

Conclusions: Group sequential and seamless designs allow for early termination based on futility criteria. This talk illustrates the need for statistical methods that incorporate this design feature and presents a Bayesian statistical framework to obtain well-calibrated posterior predictions for parameters of interest.

O5. Opportunities to speed up in-vitro diagnostic adoption and patient access in the UK: the pre-eclampsia testing timeline

Katie Scandrett[1], Joy Allen, Jon Deeks[1], Julia Eades, Ashton Harper, Christopher Hyde, Yemisi Takwoingi[1], David Wells
[1]*University Of Birmingham*

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

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

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

O6. Performance evaluation of predictive AI models to support medical decisions: overview and guidance

Ben Van Calster[1], Gary Collins, Andrew Vickers, Laure Wynants, Kathleen Kerr, Lasai Barreñada, Gael Varoquaux, Karandeep Singh, Karel Moons, Tina Hernandez-Boussard, Dirk Timmerman, David McLernon, Maarten van Smeden, Ewout Steyerberg

[1]*KU Leuven*

Background: The literature contains a plethora of measures to assess performance of predictive artificial intelligence (AI) models.

Objective: We aim to assess the merits of 32 key classic and contemporary performance measures when validating models with a binary outcome that are intended to be used in medical practice to support clinical decision-making.

Methods: The performance measures and accompanying plots cover five performance domains (discrimination, calibration, overall, classification, and clinical utility). The first four domains cover statistical performance, the fifth domain covers decision-analytic performance. We outline two key characteristics when selecting performance measures: (1) properness (whether the measure's expected value is optimized when it is calculated using the correct probabilities), and (2) focus (whether the measure reflects either a purely statistical aspect of performance, or reflects decision-analytic performance by properly considering misclassification costs).

Results: Fifteen measures violate one or both characteristics, such that we warn against their use. Interestingly, all classification measures (such as classification accuracy and F1) are improper for a clinically relevant decision threshold t.

Conclusions: We recommend to always report the following core set of measures/plots: a risk distribution plot per outcome category, AUROC, calibration plot, and a clinical utility measure such as net benefit with a decision curve.

O7. A software implementation for sample size calculation targeting precise risk predictions

Joie Ensor[1,2], Rebecca Whittle[1,2], Richard Riley[1,2]

[1]Department of Applied Health Sciences, School of Health Sciences, College of Medicine and Health, University of Birmingham, [2]National Institute for Health and Care Research (NIHR) Birmingham Biomedical Research Centre

Background: Clinical prediction models increasingly inform individual patient care, yet many are developed using inadequate sample sizes, leading to unstable risk predictions. While existing sample size criteria focus on minimising overfitting and ensuring overall model performance, they may not guarantee acceptably precise individual risk predictions. This disconnect between sample size methodology and the core purpose of prediction models – making reliable individual predictions – represents a critical gap in current practice.

Objectives: To demonstrate the practical implementation of a novel sample size methodology that targets precise individual risk estimates, through our newly developed software package pmstabilityss, available in R, Python, and Stata.

Methods: The pmstabilityss package implements an innovative five-step process to determine required sample size based on prediction uncertainty. Unlike traditional approaches, our software quantifies and visualises how sample size impacts the stability of individual risk predictions. Through a decomposition of Fisher's information matrix, we directly link sample size to prediction uncertainty. The software allows users to: (1) specify expected model characteristics, (2) define acceptable prediction uncertainty thresholds for different risk ranges, (3) visualise uncertainty-sample size relationships through instability plots, and (4) examine prediction stability across clinically relevant subgroups.

Results: Through worked examples, we demonstrate how pmstabilityss reveals that traditionally recommended sample sizes often yield an unacceptably wide range of individual risk predictions. Our software enables researchers to determine the sample size needed for stable individual predictions prospectively, transforming the focus from overall model metrics to reliable individual predictions. The software's instability plots provide an intuitive tool for communicating sample size requirements to stakeholders, particularly when precision requirements vary across the risk spectrum.

Conclusions: The pmstabilityss package represents a paradigm shift in prediction model development, moving beyond traditional sample size calculations to directly address individual risk prediction stability. This novel approach better aligns study design with the goal of reliable individual risk prediction, potentially improving both the utility and fairness of clinical prediction models.

O8. Developing a clinical prediction model with a continuous outcome: sample size calculations to target precise predictions

Rebecca Whittle[1,2], Richard D. Riley[1,2], Lucinda Archer[1,2], Gary S. Collins[3], Paula Dhiman[3], Amardeep Legha[1,2], Kym I.E. Snell[1,2], Joie Ensor[1,2]

[1]Department of Applied Health Sciences, School of Health Sciences, College of Medicine and Health, University of Birmingham, [2]National Institute for Health and Care Research (NIHR) Birmingham Biomedical Research Centre, [3]Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford

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

Objectives: We propose a new approach for calculating the sample size required to target precise individual-level predictions when developing a prediction model for a continuous outcome.

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

Results: The approach allows researchers to examine anticipated interval widths of individual predictions based on one particular sample size (i.e., of a known existing dataset), or to identify the sample size needed for a new study aiming to target a certain level of precision (e.g., a new cohort study). Additionally, this can be examined in particular subgroups of patients to help improve fairness of the model. We use a real example predicting Forced Expiratory Volume (FEV) in children to showcase how the approach allows researchers to calculate and examine expected individual-level uncertainty interval widths for particular sample sizes. We also showcase our new software module pmstabilityss.

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

O9. Adapting sample size calculations for the development of prediction models to control for model stability

Menelaos Pavlou[1], Gareth Ambler[1]
[1]*Department of Statistical Science, University College London*

Background: Use of recently proposed sample size calculations when developing risk models can lead to more reliable models. They calculate the sample size to ensure that the expected calibration slope (CS), as a measure of model overfitting, will meet a target value (commonly 0.9) when Maximum Likelihood Estimation (MLE) is used. A perfectly calibrated model has CS=1. These calculations require information on the number of predictors (p) and anticipated outcome prevalence and c-statistic.

Objectives: In practice, CS will vary over repeated samples of the recommended size. An aspect of model performance not currently considered is model stability, effectively reflected by the variability in CS. The sample size to ensure good performance on average as well as model stability is what we are concerned with in this work.

Methods: We use simulation to investigate model stability when varying p, c-statistic and prevalence. Model stability can be quantified, for example, by the Probability of obtaining a model with Acceptable Calibration (PAC), defined here as CS [0.85-1.15]. We propose an adaptation of the sample size calculations to control model stability, by ensuring that PAC is sufficiently high. In the same simulation we also explore the performance of a simple uniform shrinkage approach (using bootstrapping).

Results: When adhering to the existing sample size recommendations, the variability in the CS increased substantially with decreasing p (although CS=0.9 was met on average). Consequently, PAC was often low, particularly for p<10. Our proposed adaptation resulted in higher sample sizes than the existing recommendations for p<15, and similar sizes for higher p. Applying uniform shrinkage led to substantially higher PAC unless the number of predictors was very small (≤6).

Conclusions: Sample size calculations for the development of prediction models must cost control for model stability, in addition to average performance. Post-estimation shrinkage can be beneficial unless p is very small.

References:
1. Riley R.D. et al. Calculating the sample size required for developing a clinical prediction model (2020) BMJ; doi:10.1136/bmj.m441
2. Pavlou M. et al. An evaluation of sample size requirements for developing risk prediction models with binary outcomes (2024) BMC Med Res Methodol (2024). https://doi.org/10.1186/s12874-024-02268-5

O10. How to Handle Missing Data across the Development, Validation and Implementation of Clinical Prediction Models

Antonia Tsvetanova[1], Matthew Sperrin[1], David Jenkins[1], Niels Peek[2], Iain Buchan[3], Marcus Taylor[4], Angela Wood[5,6], **Glen Martin**[1]

[1]Division of Informatics, Imaging and Data Science, Faculty of Biology, Medicine and Health, University of Manchester, [2]The Healthcare Improvement Studies Institute (THIS Institute), Department of Public Health and Primary Care, University of Cambridge, [3]Civic Health Innovation Labs, The University of Liverpool, [4]Department of Cardiothoracic Surgery, Manchester University Hospital NHS foundation Trust, [5]British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, Victor Phillip Dahdaleh Heart and Lung Research Institute, University of Cambridge, [6]Health Data Research UK

Background: Missing data is a significant challenge in the clinical prediction model (CPM) pipeline, yet guidance on how to handle missing data across development, validation and deployment stages is limited. Ultimately, we want the imputation approach during model development and validation to provide an unbiased estimate of the CPM's predictive performance when the CPM is deployed within that target population. We call this 'compatibility'.

Objectives: To identify compatible combinations of imputation methods under varying missingness assumptions, assuming validation data reflect the same population as the model will be deployed in.

Methods: A simulation study and an empirical study using thoracic surgery data compared CPM performance across combinations of the following imputation methods across the pipeline: complete case analysis, mean/mode imputation, single deterministic regression imputation, multiple imputation, and pattern sub-model. A combination was compatible if the estimated predictive performance upon validation using a given imputation method was unbiased with respect to the performance once deployed under a given, potentially different, imputation method.

Results: For CPMs deployed requiring all data to be available, then multiple imputation during development and validation were compatible if missing data were MCAR or MAR; if the data were MNAR, then only developing and validating on data with no missing data was compatible. For CPMs deployed allowing for missing data, only imputing data in development and validation using the same missing data method planned during deployment was compatible. Commonly used combinations of imputation methods across the CPM pipeline result in biased predictive performance estimates.

Conclusions: The choice of how to handle missing data in CPM development and validation should be based on how missing data will be handled when the model is implemented. If planning to deploy a CPM without allowance for missing data, then CPM production should occur on complete data or multiple imputed datasets. If planning to deploy the CPM allowing for missing data, then CPM production should use the same imputation strategy as that planned to be used at deployment. These findings emphasize the importance of consistent missing data handling across the CPM pipeline to ensure unbiased and reliable predictive performance.

O11. Simulation Study Examining Impact of Study Design Factors on Variability Measures

Laura Quinn[1,2], Jon Deeks[1,2], Yemisi Takwoingi[1,2], Alice Sitch[1,2]
[1]University of Birmingham, [2]Birmingham Biomedical Research Centre

Introduction: Interobserver variability studies in diagnostic imaging are crucial for assessing the reliability of imaging test interpretations between different observers. The design and conduct of these studies are influenced by various factors that can impact the calculation and interpretation of variability estimates. These factors include participant sample size, condition prevalence, diagnostic test discrimination, and reader error levels.

Methods: Data was simulated for a study design with binary outcomes and two interpretations for each patient. A range of scenarios were simulated, varying participant sample size (25 to 200), condition prevalence (5% to 95%), diagnostic test discrimination (good, reasonable, poor), and reader error levels (low, medium, high). For each combination, 1,000 simulations were performed, and variability measures (percentage agreement, Cohen's kappa, Prevalence-Adjusted Bias-Adjusted Kappa (PABAK), Krippendorff's alpha, and Gwet's AC coefficient) were calculated, along with sensitivity and specificity.

Results: The study showed that increased sample size consistently produced more precise variability estimates across all scenarios. Percentage agreement consistently showed the highest values among the variability measures. PABAK and Gwet's AC coefficient demonstrated greater stability and less sensitivity to condition prevalence compared to Cohen's kappa and Krippendorff's alpha, which showed more variable performance. As diagnostic test discrimination decreased and reader error increased, all variability measures showed a decline.

Conclusion: These findings show the importance of considering different factors in assessing interobserver variability in diagnostic imaging tests. Different variability measures are affected in distinct ways by participant sample size, condition prevalence, diagnostic test discrimination, and reader error levels. By providing guidance on designing interobserver variability studies, future studies can be improved, providing more accurate information on the reliability of diagnostic imaging tests, leading to better patient care.

O12. Real world implementation of the Biomarker Toolkit: a Tool aiming to quantifiably assess biomarker utility and guide development

Katerina-Vanessa Savva[1], Alice Baggaley[1], Silvana Debernardi[2], Tatjana Crnogorac-Jurcevic[2], Melody Ni Zhifang[1], George B. Hanna[1], Christopher Peters[1]
[1]Imperial College London, [2]Barts Cancer Institute

Background: Increased resources have been spent on cancer biomarker discovery, for both prognostic and diagnostic purposes, but very few of these biomarkers have been clinically adopted. To bridge the gap between biomarker discovery and clinical use, we have previously developed and validated the Biomarker toolkit (Savva et al.,2023). This tool aims assess biomarker potential and then, more importantly, guide its further development. This study applies the Biomarker Toolkit to early-phase cancer biomarkers in collaboration with CRUK Horizons and Pancreatic Cancer Group(PCG), at Barts Cancer Institute. It aims to identify research gaps, guide biomarker development, and evaluate the tool's real-world impact and usability.

Methods: The Biomarker toolkit was developed using mixed methodology, including systematic literature searches, semi-structured interviews, and a two-stage Delphi Survey. Validation involved systematic reviews related to successfully clinically implemented and stalled of breast and colorectal cancer biomarkers, with aggregated scores assigned based on checklist criteria presence. We aim to apply the toolkit to biomarkers at various development stages using collaborator databases, including CRUK Horizon and PCG. Data on selected biomarkers will be gathered, scored, and evaluated for impact and utility through stakeholder interviews and UMUX metrics.

Results: The PCG developed a urine-based biomarker panel and PancRISK for early pancreatic cancer detection. Using systematic literature searches (Medline, Embase) and internal reports, we applied the Biomarker Toolkit (Savva et al., 2023). The biomarker scored 40.25% for clinical validity (reference score: successful biomarkers:41.51%, stalled biomarkers:36.47%), 49.35% for analytical validity (successful biomarkers:49.35%, stalled biomarkers:46.42%), and 9.62% for clinical utility (successful biomarkers:54.16%, stalled biomarkers:15.82%). Key gaps identified included Human Factor analysis and budgetary impact assessment. The PCG found the report valuable, guiding collaboration with the NIHR HealthTech Research Centre – in vitro diagnostics to address research gaps. Results from CRUK Horizon biomarkers are pending.

Conclusion: This study applied the Biomarker Toolkit to real-world early detection biomarkers, identifying research gaps at any stage of development to guide their trajectory toward clinical utility. The toolkit helps to: i) prioritise biomarkers with high clinical implementation potential, ii) shape study design, and iii) provide a framework to inform impactful research. By supporting diagnostic biomarker translation, it reduces excessive discovery costs and promotes early patient diagnosis.

O13. Methodology to create evidence-based testing panels for monitoring long-term conditions in primary care

Martha Elwenspoek[1], Rachel O'Donnell[1], Alice Malpass[1], Katie Charlwood[1], Mary Ward[1], Howard Thom[1], Jonathan Banks[1], Clare Thomas[1], Hayley Jones[1], Jonathan Sterne[1], Lewis Buss[1], Francesco Palma, Christina Stokes, Alastair Hay[1], Jessica Watson[1], Penny Whiting[1]
[1]*University of Bristol*

Background: Patients with long term conditions (LTC) have regular monitoring appointments including blood tests. These tests aim to monitor disease progression, treatment response, and detection of complications. The evidence base for current testing recommendations is weak because measuring patient benefits or harms of regular monitoring is challenging and are dependent on what is done in response to the test result.

Objectives: To develop a methodology for creating evidence-based testing strategies to monitor people with LTCs.

Methods: We identified a list of commonly used blood tests. We defined a series of filtering questions to determine whether there was evidence to support the rational of monitoring, such as 'can the GP do anything in response to an abnormal test result?'. Through a series of rapid reviews we identified evidence to answer each question. At consensus meetings clinicians and patients voted on each test for inclusion, exclusion, or further evidence. A process evaluation was performed alongside this. Additional evidence was collected by performing further rapid reviews and by analysing routinely collected healthcare data, including incidence analyses and emulating RCTs.

Results: We tested this methodology on three common LTCs: chronic kidney disease (CKD), type 2 diabetes mellitus (T2DM), and hypertension. We identified 18-21 commonly used tests per condition. We found that for the majority of blood tests the evidence-base was weak or absent. Only 1-2 tests per condition could be included (e.g. HbA1c and eGFR for monitoring T2DM patients) and 8-13 tests could be excluded (e.g. inflammation markers for all three conditions) based on evidence alone. The consensus group selected 4-8 tests per condition where additional evidence was needed to make a decision. The new evidence was generally weak and a final consensus meeting was necessary to finalise the testing panels based on expert opinion.

Conclusions: We are currently testing the clinical and cost-effectiveness of the developed testing panels in a cluster RCT. This methodology may be used to optimise disease monitoring of other chronic conditions. Implementing evidence-based testing panels may address some of the unwarranted variation in testing and improve patient outcomes by reducing patient harm related to over- and undertesting.

O14. Measurement Error: Unlocking Estimates of Test Variability From Routine Data. Methods for Statistical Analysis and a Case-Study Series

Simon Baldwin[1,2,3], Susan Mollan[3,4], Balazs Baranyi[3,4], Alice Sitch[1,2], Jonathan J Deeks[1,2]
[1]University of Birmingham, [2]NIHR Birmingham Biomedical Research Centre, [3]University Hospitals Birmingham, [4]INSIGHT Health Data Research Hub For Eye Health

Background: To consider tests for clinical application, early-stage evaluation includes estimating the measurement error (test measurement variability around the conceptual 'truth'). Whilst prospective biological variability studies (BVS') provide the ideal approach (repeated biomarker measurements taken on individuals at stated follow-up time-points, including duplicate measurements), limited funding and repeat testing burden means that they are not always feasible. Routine datasets are a convenient, inexpensive source of repeated biomarker measurements, but methodological assumptions for analysing BVS' do not hold for data collected at irregular time-points, or without duplicate measurements.

Objective: Investigate three methods identified from scoping test variability literature, for estimating measurement error from routinely collected biomarker data: a linear mixed effects (LME) approach [1]; baseline-pairs approach [2]; and autocorrelation approach [3].

Methods: Assess how estimates generated by the three methods were affected by clinical context in a case-study series: blood pressure variability in children (CS1); serum albumin variability in adults with progressive cholangitic disease (CS2); and retinal nerve thickness variability (RNFL_G, μm^2) in patients with stable ocular hypertension (CS3). Measurement error (and within-individual variability) estimates were generated for each model.

Results: CS1: measurement error estimates were similar across the three methods; the within-individual variability in BP was >4x smaller than the measurement error. CS2: convergence issues affected the autocorrelation model, due to the constant-variation assumption underlying the method and the increasing-variation in the data (over time). For the LME and baseline-pairs models, there were no convergence issues in CS2 (both methods assume increasing-variation over time); however, measurement error estimates were increased by ~30% in deceased patients, compared to survivors. CS3: measurement error estimates were similar for the LME and baseline-pairs models (~8μm^2), but ~2x smaller for the autocorrelation model (~4μm^2); however, unlike CS1, the within-individual variability in RNFL_G was >3x larger than the measurement error.

Conclusions: There were important differences in the estimates generated by the three methods, dependent on clinical scenario. Measurement error was shown to increase with disease progression, and should therefore be assessed in steady-state individuals. However, to investigate which other data characteristics impact on the choice of method used to estimate measurement error, simulation studies were necessary.

O15. Evaluation of diagnostic tests with spatially or temporally clustered data, part 1: The choice of estimands and estimators affects results and interpretation

Nicole Rübsamen[1], Julia Böhnke[1], André Karch[1], ELISE Study Group[2], Philipp Weber[3], Antonia Zapf[3]

[1]Institute of Epidemiology and Social Medicine, University of Münster, [2]ELISE project, [3]Institute of Medical Biometry and Epidemiology, University Medical Center Hamburg-Eppendorf

Background: If an index test is applied multiple times (leading to spatially or temporally clustered data), estimation of diagnostic accuracy must account for intra-patient dependencies to avoid bias. Currently, recommendations for handling these dependencies are not available. Several approaches were proposed for the analysis of spatially clustered data, which can be roughly divided into parametric (random effect models), semi-parametric (GEEs), and non-parametric approaches. We showed that using temporally clustered data requires researchers to consider methodological choices and predefined estimands [1].

Objectives: To illustrate how different methodological choices for diagnostic accuracy involving temporally clustered data affect results and interpretation.

Methods: We applied three different methods [2–4] to evaluate the diagnostic accuracy of a knowledge-based decision model (the index test that is applied every minute) on a systemic inflammatory response syndrome dataset from the ELISE study [1] versus clinicians' diagnoses as reference standard. We used two different levels as proposed in [1]: the block-level and the time-level with hours as time units.

Results: Intra-method comparison highlights that depending on the level the point estimates and their two-sided 95% confidence intervals differ widely (Fig. 1). However, inter-method comparison depicts that the point estimates of the same level across the methods vary little while their 95% confidence intervals vary considerably, which may lead to different conclusions in confirmatory trials.
Conclusions: We lack a comprehensive framework for the planning and analysis of diagnostic studies with spatially or temporally clustered data, and of a translation of this framework into research practice. With our project ClusterDiag, we want to enable researchers to define the appropriate estimand for their study with clustered data, to choose the optimal approach as an estimator, and to obtain valid estimates.

Funding: German Research Foundation [539658720 "ClusterDiag"]; German Federal Ministry of Health [2520DAT66A "ELISE"].

Remark: This presentation should be given together with "Evaluation of diagnostic tests with spatially or temporally clustered data, part 2: Scoping review of different methods for estimating diagnostic accuracy for clustered data". For illustration purposes, we only compare three methods (parametric, semi-parametric, and non-parametric) from this scoping review.

References:
[1] doi.org/10.1016/j.jclinepi.2024.111314
[2] doi.org/10.1002/(SICI)1097-0258(19970615)16:11<1263::AID-SIM550>3.0.CO;2-M
[3] doi.org/10.1002/sim.688
[4] doi.org/10.1016/j.csda.2011.03.022

O16. Value-of-Information Analysis for External Validation of Risk Prediction Models in Multicenter Studies and Systematic Reviews

Laure Wynants[1], Sabine Grimm[1], Andrea Gabrio[1], Ben Van Calster[2], Ewout Steyerberg[3], Andrew Vickers[4], Mohsen Sadatsafavi[5]

[1]Maastricht University, [2]KU Leuven, [3]University Medical Center Utrecht, [4]Memorial Sloan Kettering Cancer Center, [5]University of British Columbia

Background: Clinical prediction models should demonstrate clinical utility before they are used for medical decision making. The Net Benefit (NB) framework can be used to quantify clinical utility, and the expected value of perfect information (EVPI) was recently proposed to quantify the consequence of uncertainty regarding which strategy (use model, treat all, treat none) is superior. Put simply, the expected value of perfect information is equivalent to the expected loss due to uncertainty: the probability of choosing the wrong strategy, multiplied by the average consequence of being wrong (NB loss). EVPI is currently focused on single validation studies, but variance in utility across local settings (e.g., hospitals, countries) can be investigated in multicenter validations and systematic reviews.

Objective: To propose value-of-information (VoI) measures that build on random effect meta-analysis to quantify the expected gain from eliminating uncertainty regarding the superior strategy (model, treat all, treat none), locally and globally.

Methods: We distinguish between perfect and partial perfect information, at the global and local level, upon which different types of EVPI measures are defined. The resulting EVPI concepts are illustrated using data from a systematic review of the predictive performance of a diagnostic model (ADNEX) for ovarian cancer.

Results: Based on validation data from 37 centers, the ADNEX model was globally superior to default strategies, with no expected NB gain from further validation studies to eliminate uncertainty (EVPI global =0). However, we estimated that in 4% of centers another strategy would be superior. The corresponding expected gain from eliminating uncertainty regarding the locally superior strategy (EPVI local) is 1,332 net avoided false positive classifications per year for Europe. The expected gain from determining the best local strategy conditional on known prevalence, but not on local performance (expected value of partial perfect information), is a net 153 avoided false positive classifications per year.

Conclusions: The Net Benefit of a model is likely to vary between settings, as might the optimal strategy. In multicenter validation studies and meta-analyses, VoI metrics can be defined globally and locally. The methods proposed here can be used in practice to determine whether more local research is needed.

O17. Comparing Performance of Methods that Correct for Data Distribution Shift when Developing Clinical Prediction Models: A Simulation Study

Haya Elayan[1]
[1]*University Of Manchester*

Background: Clinical Prediction Models (CPMs) compute the risk of a diagnostic or prognostic outcome, given a set of predictors, but are developed assuming that the data distribution remains constant throughout the CPM pipeline. However, in dynamic healthcare environments, Data Distribution Shift can occur during CPM development, where the underlying distribution of the data changes over time, or space. This shift can result in a mismatch between the development population (Source population) and deployment population (Target population), potentially degrading model performance due to the inclusion of data samples from both populations in the development dataset. There is opportunity to exploit or adjust for shift within the development dataset to address the shift between the development and deployment phases, without requiring additional samples from the target population. This study aims to compare the predictive performance of CPM development and updating methods to account for data distribution shift in a development dataset sampled from both source and target populations.

Methods: We perform simulations to investigate various distribution shifts. We developed multiple CPMs, and compare their performance using calibration, discrimination, and prediction stability metrics. Our simulation process includes generating source and target populations, building models on a development dataset that combines samples from source and target populations, and validating these models on the remaining target population. Comparisons include naive logistic regression, regression updating (recalibration and Bayesian updating), and importance weighting using propensity scores.

Results and Conclusion: Logistic Recalibration and Intercept Recalibration improve performance over naïve regression. However, the former showed greater variability in calibration slopes and more instability in calibration curves in all scenarios compared to other models, while the latter performed worse in calibration slope under predictor-outcome association shift compared to other models. Bayesian updating also resulted in misscalibration-in-the-large for event rate and predictor-outcome association shifts. Additionally, using a subset of the data that is most reflective to the target population for model development was the least effective method due to the smaller sample size. In contrast, the weighting method using propensity scores showed consistent results with improved performance to other methods for many scenarios, offering a promising alternative to the compared methods.

O18. Use of statistical process control to monitor calibration-in-the-large of a clinical prediction model

David Jenkins, Glen Martin[1], Niels Peek[2], Mamas Mamas[3], Matthew Sperrin[1]
[1]Faculty of Biology, Medicine and Health, University of Manchester, [2]The Healthcare Improvement Studies Institute (THIS Institute), University of Cambridge, [3]Keele University

Background: Clinical prediction models (CPMs) can be updated with contemporary data to address calibration drift but it is often unclear if and when this should happen. This study aimed to explore the use of statistical process control (SPC) to identify if and when a CPM should be updated.

Methods: Our SPC method monitors the difference between predicted and observed risks (calibration-in-the-large), and generates an alert when there is evidence of systematic over- or under-prediction. We performed a simulation study based on an existing CPM for 30-day mortality after percutaneous coronary intervention, generating outcomes under varying degrees of miscalibration. Time of an alert, generated from using 3 and 4 standard deviations (SDs) as the SPC control limits, and the number of individuals that would have incorrectly been classified using a risk threshold were recorded.

Results: For data generating scenarios where the average systematic over prediction was 1.2-fold the true risk, an alert was triggered on average after 7510 patients (95% CI: 294, 27136) and 16022 patients (95% CI: 1863, 42160) for the 3SD and 4SD control limits, respectively. For scenarios where the average overestimation of risk was 4.7-fold an individual's true risk, the average time to alert for the 3 and 4SD control limits were 246 (95% CI: 175, 320) and 246 (95% CI: 175, 321), respectively. When there was no miscalibration, 10% of the iterations resulted in an incorrect alert for the 3SD control limit, compared to 1.2% using the 4SD control limit. For all risk thresholds, the 4SD and 3SD control limit incorrectly classified up to 17 and 5 persons before an alert, respectively. Conversely, no control limit resulted in up to 6000 misclassifications.

Conclusion: SPC can be used as a way of continually monitoring the calibration-in-the-large of a CPM, to suggest when a model should be recalibrated. Control limits could be chosen based upon the clinical scenario, healthcare setting and acceptable error rates.

O19. Combining calibration plots from multiple centers or datasets

Lasai Barreñada[1,3], Laure Wynants[1,2,3], Ben Van Calster[1,3]
[1]Department of Development and Regeneration, KU Leuven, [2]Department of Epidemiology, CAPHRI Care and Public Health Research Institute, Maastricht University, [3]Leuven Unit for Health Technology Assessment Research (LUHTAR), KU Leuven

Background: Evaluation of clinical prediction models in multiple centers or datasets (clusters) is becoming more and more common. A comprehensive evaluation includes an assessment of the agreement between the predicted risks and the observed outcomes, also known as calibration. Calibration often varies importantly between clusters.

Objectives: We present three novel random effects pooling methods to obtain flexible calibration plots with prediction intervals, and compare these methods to cluster-ignorant plots.

Methods: First, clustered group calibration (CG-C) groups cluster-specific observations in quantiles and pools each cluster's average risks and prevalence with a bivariate random effects model. Second, two stage meta-analysis (2MA-C) first trains center-specific flexible curves using splines or LOESS and then pools the individual observed proportion across the predicted risks. Third, mixed model calibration (MIX-C) fits a multilevel flexible model with splines and random intercepts and slopes by cluster. We compared the methods in a case study on ovarian tumor malignancy (n=2,489), a simulation study (AUC 0.75 or 0.90, ICC 0.05 or 0.2, 5 or 30 clusters, 20 or 200 events per variable in each cluster,), and a synthetic clinical data study (n=10,000,000). We evaluate the mean squared calibration error (MSCE) that measures the difference between the true and observed calibration curve. In addition we calculate coverage of 95% prediction intervals.

Results: Across simulation scenarios the best median MSCE was for MIX-C and 2MA-C with splines. However, in the synthetic data analysis 2MA-C with splines performed worse than 2MA-C with LOESS but MIX-C worked consistently well. CG-C with 10 quantiles also worked well but depends heavily on the number of quantiles. Cluster-ignorant flexible calibration worked competitively when sample size is limited but does not provide prediction intervals. MIX-C had lowest MSCE and provided close to nominal coverage of 95% prediction intervals for hypothetical new centers (94% coverage).

Conclusions: MIX-C showed to be the best approach overall to present an average curve for calibration with sensible prediction intervals that represent the calibration of hypothetical new centers.

O20. Network meta-analysis of prediction models using aggregate or individual participant data - A scoping review and recommendations for reporting and conduct

Maerziya Yusufujiang[1], Johanna A.A.G. Damen[1], Demy L. Idema[1], Ewoud Schuit[1], Karel G.M. Moons[1], Valentijn M.T. de Jong[1,2]
[1]Julius Center for Health Sciences and Primary Care, UMC Utrecht, Utrecht University, [2]Data Analytics and Methods Task Force, European Medicines Agency

Background: Prediction models are essential in clinical decision-making for estimating risks of current (diagnosis) or future (prognosis) outcomes. Network meta-analysis (NMA) serves as a powerful tool to compare the performance of multiple prediction models simultaneously. There is limited knowledge on methodology and reporting of studies employing NMA for prediction model evaluation.

Objective: To provide an overview of NMAs of prediction model evaluations, regardless of whether they used aggregate data (AD) or individual participant data (IPD).

Methods: We searched PubMed up to 2nd May 2024 to identify studies that addressed the evaluation of prediction models performance using NMA. We included articles within the medical domain that employed NMA to compare and assess at least three diagnostic or prognostic prediction models. We summarized the identified studies based on their application, data type, clinical contexts, evaluation metrics applied. Additionally, we examined the statistical approaches employed, the NMA assumptions, and the ranking methods used for model comparison.

Results: After screening 1,795 articles, 17 were included. Fifteen studies (88.2%) used AD, while two (11.8%) utilized IPD. Hospital care was the most common setting (n = 12; 70.6%), with respirology (n = 5; 29.4%) and cardiology (n = 4; 23.5%) as the most common clinical domains.
NMA assumptions were addressed differently across NMA: 11.8% (n = 2) discussed transitivity, similarity, or exchangeability, and 58.1% (n = 10) tested for consistency. The statistical approach varied, with 41.2% of studies employing Bayesian and 29.4% using frequentist approaches. SUCRA was the predominant ranking method (n = 12). Most studies included 5–10 models in the network, with two studies analyzing more than 20.

Performance metrics studied varied, with 56.3% of studies (n = 9) reporting discrimination measures, such as c-statistics, while none reported calibration. Sensitivity or specificity was provided in 64.7% of studies (n = 11).

Conclusions: This review highlights the limited but diverse use of NMA in evaluating prediction models, with a predominant reliance on aggregate data and inconsistent consideration of key assumptions and performance metrics. Improvement of reporting of NMA is needed and future research should focus on advancing NMA methods to improve prediction model validation and reporting.

O21. Evaluating Diagnostic Tests Against Composite Reference Standards: Quantifying and Adjusting for Bias

Vera Hudak[1], Nicky J. Welton[1], Efthymia Derezea[1], Hayley Jones[1]
[1]*University Of Bristol*

Background: Composite reference standards (CRSs) are often used in diagnostic accuracy studies in situations where gold standards are unavailable or impractical to carry out on everyone. Here, the test under evaluation is compared with some combination (composite) of results from other tests. We consider a special case of CRS, which we refer to as a 'check the negatives' design. Here, all study participants receive an imperfect reference standard, and those who test negative on this are additionally tested with the gold standard. Unless the imperfect reference standard is 100% specific, some bias can be anticipated.

Methods: We derive algebraic expressions for the bias in the estimated accuracy of the test under evaluation in a 'check the negatives' study, under the assumption that test errors are independent given the true disease status. We then describe how this bias could be adjusted for using a Bayesian model, with an informative prior for the specificity of the imperfect reference standard based on external information. This proposed adjustment method is evaluated through a simulation study, under varying combinations of test accuracy, study sample size and disease prevalence, under the assumption that the informative prior distribution for specificity is correctly centred around its true value. The impact of increasing uncertainty in this prior distribution is also evaluated.

Results/Conclusions: In a 'check the negatives' study, under the assumption of conditional independence of errors made by the test under evaluation and the imperfect reference standard, the estimated specificity is unbiased but the sensitivity is underestimated. Preliminary findings suggest that the Bayesian model will always reduce bias and can successfully eliminate it in some, but not all, scenarios. Full simulation results and their implications will be presented at the conference.

Keywords: Composite reference standard, Bayesian models, simulation study, bias-adjustment

O22. Improving the reference standard in diagnostic accuracy studies: Evaluating a latent class model against a panel of expert clinicians

Nandini Dendukuri[2], **Tom Parry**[1], Sue Mallett[1], Steve Halligan[1]
[1]University College London, [2]McGill University

Background: Without a perfectly accurate test, we often rely on panel of expert clinicians to set the reference standard. The panel classifies participants using multiple imperfect tests but is inefficient and prone to bias from interpersonal dynamics. Latent class models (LCMs) are a promising alternative, offering more efficient and objective classification using the same imperfect tests. However, the lack of evaluations of LCMs against panels has limited broader adoption.

Objectives: To conduct the first evaluation of an LCM against a panel of expert clinicians as the reference standard in an existing diagnostic accuracy trial.

Methods: We analysed data from participants prospectively and consecutively recruited with high suspicion of active TICD across eight NHS hospitals (METRIC [ISRCTN03982913]). All participants underwent MRI and US as index tests. The panel classified disease presence using colonoscopy, histology, C-reactive protein, or faecal calprotectin when available. We collaborated with expert clinicians to prespecify tests for our LCM, including MRI, US, the tests used by the panel, and Harvey-Bradshaw index. We fit a 2-class Bayesian LCM with flat priors and a random effect to account for conditional dependence. We then sorted the individual participant disease probabilities estimated by our LCM and applied a cutoff to achieve the prevalence estimated by our LCM.

Results: In our analysis of 284 participants, the panel classified 69% (95% CI 63, 74) of participants as disease positive, while our LCM classified 71% (95% CrI 61, 78). Relative to the panel, our LCM had an 88% (95% CI 84, 91) probability of agreeing in disease-positive participants and a 72% (95% CI 66, 77) probability of agreeing in disease-negative participants. If the panel could have used MRE and US without adding incorporation bias, we believe the negative-specific agreement would have improved. This is because our LCM considered MRE and US highly accurate tests.

Conclusions: Our findings demonstrate the potential of LCMs as an efficient and viable method to set the reference standard, allowing the inclusion of highly accurate index tests without adding bias. Future analyses will evaluate LCMs against panels for other diseases and study designs.

O23. Examining the Association between Estimated Prevalence and Diagnostic Test Accuracy Using Directed Acyclic Graphs

Yang Lu[1], Robert Platt[1,2], Nandini Dendukuri[1,3]
[1]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, [2]Department of Pediatrics, McGill University, Montreal, QC, Canada, [3]Department of Medicine, McGill University, Montreal, QC, Canada

Background: There have been reports of correlation between estimates of prevalence and test accuracy across studies included in diagnostic meta-analyses. The direction of the correlation could be positive or negative. It has been hypothesized that the apparent association arises because of certain biases commonly found in diagnostic accuracy studies. A theoretical explanation for this hypothesis has not been studied systematically.

Methods: In this manuscript, we employ directed acyclic graphs (DAGs) to illustrate common biases in DTA studies and to define the resulting data-generating mechanism behind a diagnostic test accuracy (DTA) meta-analysis. Using simulation studies covering a range of scenarios we examine how these common biases can produce a correlation between estimates of prevalence and test accuracy, and what factors influence its magnitude and direction.

Results: We found that a spurious association arises in the absence of a perfect reference test while a genuine association arises in the presence of a covariate that simultaneously causes spectrum effect and is associated with the prevalence. We also show the spurious association can be removed when correcting reference standard bias using latent class meta-analysis.

Conclusions: As part of the risk of bias evaluation in DTA meta-analyses, an observed association between estimates of prevalence and accuracy should be explored to understand if it is genuine or spurious, and steps taken to adjust for latent or observed variables if possible.

O24. Diagnostic accuracy of tests for SARS-CoV-2 acute infection: Distinguishing measurands from target conditions

Joanna Merckx[1,2], Ian Schiller[2], Yap Boum[3], Patrick M Bossuyt[4], Nandini Dendukuri[2]
[1]Mcgill University, Department of Epidemiology, Biostatistics and Occupational Health, [2]Research Institute of the McGill University Health Centre, [3]Epicentre, Public Health Emergency Operation Center, Ministry of Public Health, [4]Department of Epidemiology & Data Science, Amsterdam Public Health, Amsterdam University Medical Center

Background: Test accuracy evaluation for SARS-CoV-2 infection is complicated by the lack of a perfect reference. Additionally, tests have different measurands further challenging performance estimation. Other issues are the need to deconflate the time-varying prevalence of antibodies (the measurands), and the time-invariant ability of the assay to measure antibodies when assessing sensitivities. We provide an alternative approach estimating the accuracy of PCR, antigen, and antibody tests for the diagnosis of the target conditions acute and past SARS-CoV-2 infection and their prevalence. We apply our methods to a Cameroonian cohort of 1,194 adults tested at multiple-time points.

Methods: We decompose the accuracy question into its elements: i) the tests under evaluation, ii) their measurands and iii) the target conditions. We use directed acyclic graphs (DAG) to visualize these elements. We use latent class analysis (LCA) to model the relationships in the DAG and Bayesian inference to obtain estimates of sensitivity, specificity and prevalence. We introduce two random effects to capture the dependence between the measurands due to acute and past infection. We represent the results as posterior distribution medians with 95% credible intervals (CrI) and compare with a measurand naïve LC model.

Results: We estimate the prevalence of acute and past infection as 20% (95%CrI 17; 24) and 26% (95%CrI 20; 32), respectively. The sensitivity of the measurands with regards to acute infection were 75% (IgM t2), 84% and 91% (IgG t2 and t3), while the sensitivity of the antibody test for the measurands IgM and IgG were 65% and 88%, respectively, at all three time points. The sensitivity of the antibody tests with respect to acute infection varied from 13% to 81%, while being similar (IgM test 57%, IgG test 85%) with respect to past infection over time.

Conclusions: By distinguishing the target conditions from the measurands of the observed tests, we were able to estimate that prevalence of antibodies increases over time following SARS-CoV-2 infection whereas the sensitivity of antibody tests with respect to their measurand is consistent over time. This adds nuance to previous reports that suggest the sensitivity of these tests increases over time.

O25. The estimand framework for diagnostic accuracy studies

Alexander Fierenz[1], Mouna Akacha[2], Norbert Benda[3], Mahnaz Badpa[1], Patrick M.M. Bossuyt[4], Nandini Dendukuri[5], Britta Rackow[1], **Antonia Zapf**[1]
[1]University Medical Center Hamburg-Eppendorf, [2]Novartis Pharma AG, [3]Federal Institute for Drugs and Medical Devices, [4]Amsterdam University Medical Center, [5]McGill University

Background: Diagnostic accuracy studies investigate how well an index test distinguishes between two conditions. To ensure valid conclusions on the diagnostic accuracy, all study aspects must be defined prior to study start. For therapeutic trials, the estimand framework was presented in the ICH E9 Appendix '[...] to strengthen the dialogue between disciplines involved in the formulation of clinical trial objectives, design, conduct, analysis and interpretation [...]' [1]. However, this framework is not directly applicable to diagnostic accuracy studies [2].

Objectives: Generating an estimand framework for diagnostic accuracy studies.

Methods: We identify attributes for an estimand in diagnostic accuracy studies. Furthermore, we consider different strategies to deal with interfering events (analogous to intercurrent events in therapeutic trials), which may impact test results or decisions. To illustrate the approach, we apply the framework to a hypothetical study in which the accuracy of computed tomography (CT) scanning in detecting lung carcinoma is to be determined illustrating different potential issues with respect to interfering events e.g. referring to premature termination of a scan due to dyspnoea or claustrophobia (which may or may not be informative about the disease to be diagnosed).

Results: The attributes defined for an estimand in a diagnostic accuracy study are: population, target condition, index test, accuracy measure, and interfering events. We suggest six possible strategies for dealing with interfering events. For the example study, an estimand could be: In patients with suspected lung carcinoma who fulfil the requirements for CT imaging (population), sensitivity and specificity (accuracy measures) of a CT scan (index test) will be assessed, for the detection of lung carcinoma (target condition). While premature termination of the CT scan due to dyspnea will be counted as positive result (indicative event strategy), we would use multiple imputation for termination due to claustrophobia (hypothetical strategy).

Conclusions: We recommend carefully considering and defining an estimand for confirmatory diagnostic accuracy studies. In particular, possible interfering events and strategies for dealing with them should be defined in advance and reflect the scientific question on the accuracy of the diagnostic test.

References:
1. https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e9-r1-addendum-estimands-sensitivity-analysis-clinical-trials-guideline-statistical-principles_en.pdf
2. doi.org/10.1002/pst.2395
Funding: German Research Foundation [499188607 „MisEstiDiag“]

O26. How do authors of comparative accuracy studies analyse data when reporting a comparative conclusion: methodological review?

Yaxin Chen[1], Yasaman Vali[1], Sue Mallett[2], Anne Wilhelmina Saskia Rutjes[3], Clare Davenport[4], Bada Yang[5], Mariska Leeflang[1]

[1]Department of Epidemiology and Data Science, Amsterdam UMC, University of Amsterdam, [2]Centre for Medical Imaging, University College London, [3]Saint Camillus International University of Health and Medical Sciences (UniCamillus), [4]Department of Applied Health Science, University of Birmingham, [5]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University

Background: Diagnostic accuracy studies comparing index tests (comparative accuracy studies) often report comparison results where evaluation can be biased by choices in statistical modelling. The quality assessment tool for comparative accuracy studies, QUADAS-C, currently does not include signalling questions to assess biases in the comparison due to statistical choices.

Objectives: Methodological review of statistical analyses used in comparative accuracy studies with the ultimate aim of deriving possible signaling questions for a future QUADAS-C analysis domain.

Methods: We searched all systematic reviews of diagnostic test accuracy published in 2023 using PubMed. Systematic reviews with a comparative aim and containing at least 10 comparative primary studies were selected. From all comparative studies included in these reviews, we randomly selected a subset of 200 studies. Single data extraction was conducted by five reviewers.

Results: Of 200 studies, 53 studies compared two tests, and 147 studies compared more than two tests. Eighty-three percent of the studies (166/200) drew a comparative conclusion. The three accuracy measures that were used the most for comparison were sensitivity and specificity (164/200), area under the receiver operating characteristic curve (92/200), and predictive values (52/200). About half of the studies (99/200) formally compared accuracy measures, using a statistical test. The McNemar test (22/99), DeLong's test (18/99), and the Chi-square test (7/99) were the most commonly used. Fifteen studies provided formal sample size calculations, of which nine were based on the comparative questions. Fifteen studies clearly reported missing data: five had none, six excluded patients with missing data, and four replaced missing values with zero. Ten studies (of which eight used the fully paired) took confounding into consideration and the most common method was stratification (8/10).

Conclusions: Most studies drew a comparative conclusion, but few studies reported enough transparent details about statistical methods to evaluate the comparison. These results are important to inform the risk of bias questions for the QUADAS-C analysis domain. Future steps involve the potential effects of chosen methods on final inference with respect to the comparison.

O27. PROBAST+AI: An updated quality, risk of bias and applicability assessment tool for prediction models using regression or artificial intelligence methods

Karel Moons[1], **Anneke Damen**[1], Tabea Kaul[1], Lotty Hooft[1], Constanza Andaur Navarro[1], Paula Dhiman[2], Andrew Beam[3], Ben van Calster[4,5], Leo Anthony Celi[6], Spiros Denaxas[7], Alastair Denniston[8], Marzyeh Ghassemi[9], Georg Heinze[10], André Pascal Kengne[11], Lena Maier-Hein[12], Xiaoxuan Liu[8,13,20,21], Patricia Logullo[2], Melissa McCradden[14], Nan Liu[15], Lauren Oakden-Rayner[16], Karandeep Singh[17], Daniel Ting[15,18], Laure Wynants[4,19], Bada *Yang[1], Hans Reitsma[1], Richard Riley[20,21], Gary Collins[2], Maarten van Smeden[1]*
[1]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, [2]Centre for Statistics in Medicine, UK EQUATOR Centre, Nuffield Department of Orthopaedics, Rheumatology & Musculoskeletal Sciences, University of Oxford, [3]Department of Epidemiology, Harvard T.H. Chan School of Public Health, [4]Department of Development and Regeneration, KU Leuven, [5]Leuven Unit for Health Technology Assessment Research (LUHTAR), KU Leuven, [6]Harvard Medical School, Boston, Massachusetts; Laboratory for Computational Physiology, Massachusetts Institute of Technology, [7]Institute of Health Informatics, University College London, [8]College of Medicine and Health, University of Birmingham, [9]Department of Electrical Engineering and Computer Science; Institute for Medical Engineering and Science, Massachusetts Institute of Technology, [10]Institute of Clinical Biometrics, Center for Medical Data Science, Medical University of Vienna, [11]Department of Medicine, University of Cape Town, [12]Division of Intelligent Medical Systems (IMSY), German Cancer Research Center (DKFZ), [13]University Hospitals Birmingham NHS Foundation Trust, [14]Department of Bioethics, The Hospital for Sick Children, [15]Centre for Quantitative Medicine, Duke-NUS Medical School, [16]Australian Institute for Machine Learning, University of Adelaide, [17]Department of Learning Health Sciences, University of Michigan Medical School, [18]AI Office, Singapore Health Service; Duke-NUS Medical School, [19]Department of Epidemiology, CAPHRI Care and Public Health Research Institute, Maastricht University, [20]School of Health Sciences, College of Medicine and Health, University of Birmingham, [21]National Institute for Health and Care Research (NIHR) Birmingham Biomedical Research Centre

Background: PROBAST (Prediction model Risk Of Bias Assessment Tool) is a tool to assess the risk of bias and applicability of prediction models and prediction model studies. Since its launch in 2019, there has been much progress on the methodology for prediction modelling in general and on the use of artificial intelligence (AI), including machine learning (ML) techniques, in particular.

Objectives: To develop PROBAST+AI, an updated tool for assessing quality, risk of bias, and applicability of studies developing or evaluating (validating) multivariable diagnostic and prognostic prediction models using any data analytical (prevailing statistical or AI/ML) technique.

Methods: PROBAST+AI was informed by the original PROBAST tool and user experiences, by various systematic reviews and by recent methodological developments, followed by a 3-round Delphi survey among a large, global and diverse group of key stakeholders and experts on prediction modelling, AI/ML techniques and healthcare professionals, followed by a final consensus expert meeting. Participants gave their opinion on a large series of quality, risk of bias, and applicability items, using a 5-point Likert scale. Participants were also asked to suggest new items using free-text boxes.

Results: PROBAST+AI consists of two parts: model development and model evaluation. Each part contains four domains: Participants and Data Sources, Predictors, Outcome, and Analysis. For model development, PROBAST+AI users assess the quality and applicability using 16 signalling questions. For model evaluation, PROBAST+AI users assess the risk of bias and applicability using 18 signalling questions. PROBAST+AI also has signalling questions related to prediction model fairness and algorithmic bias.

Conclusions: PROBAST+AI may replace the original PROBAST tool and provides a tool that allows all key stakeholders (e.g., prediction model developers and companies, researchers, editors, reviewers, health care professionals, patients, institutional ethical review boards, healthcare guideline developers and health policy organisations) to examine the quality, risk of bias and applicability of any type of prediction model study in the healthcare sector, regardless which data analytical (prevailing statistical or AI/ML) technique is used.

O28. Guidance for unbiased predictive information for healthcare decision-making and equity (GUIDE): considerations when race may be a prognostic factor

Keren Ladin[1], **David Kent**[2], John Cuddeback[3], O Kenrik Duru[4], Sharad Goel[5], William Harvey[6], Jinny G Park[7], Jessica K Paulus[8], Joyce Sackey[9], Richard Sharp[10], Ewout Steyerberg[11], Berk Ustun[12], David van Klaveren[13], Saul N Weingart[6]
[1]Tufts University, [2]Tufts Medical Center, [3]American Medical Group Association, [4]University of California Los Angeles, [5]Harvard University, [6]Tufts Medical Center, [7]Tufts Medical Center, [8]OM1, [9]Stanford University School Medicine, [10]Mayo Clinic Center for Individualized Medicine, [11]Leiden University Medical Centre, [12]University of California San Diego, [13]Erasmus University Medical Centre

Background: Clinical prediction models (CPMs) are tools that compute the risk of an outcome given a set of patient characteristics and are routinely used to inform patients, guide treatment decision-making, and resource allocation. Although much hope has been placed on CPMs to mitigate human biases, CPMs may potentially contribute to racial disparities in decision-making and resource allocation. While some policymakers, professional organizations, and scholars have called for eliminating race as a variable from CPMs, others raise concerns that excluding race may exacerbate healthcare disparities and this controversy remains unresolved.

Methods: The Guidance for Unbiased predictive Information for healthcare Decision-making and Equity (GUIDE) provides expert guidelines for model developers and health system administrators on the transparent use of race in CPMs and mitigation of algorithmic bias across contexts developed through a 5-round, modified Delphi process from a diverse 14-person technical expert panel (TEP).

Results: Deliberations affirmed that race is a social construct and that the goals of prediction are distinct from those of causal inference, and emphasized: the importance of decisional context (e.g., shared decision-making versus healthcare rationing); the distinction between bias and fairness; the conflicting nature of different anti-discrimination principles (e.g., anti-classification versus anti-subordination principles); and the importance of identifying and balancing trade-offs in achieving equity-related goals with race-aware versus race-unaware CPMs for conditions where racial identity is prognostically informative. The GUIDE emphasizes the consequences on decision-making of race-aware versus race-unaware prediction in terms of the harms and benefits to patients and de-emphasizes the importance of delineating the causal mechanism of the predictive effects of race.

Conclusion: The GUIDE, comprising 31 key items in the development and use of CPMs in healthcare, outlines foundational principles, and offers guidance for examining subgroup invalidity and using race as a variable in CPMs. This GUIDE presents a living document that supports appraisal and reporting of bias in CPMs to support best practice in CPM development and use.

O29. A simulation study investigating the impact of the prediction paradox on clinical prediction model performance

Samantha Pacynko[1], Matthew Sperrin[1], David Jenkins[1]
[1]*Faculty of Biology, Medicine and Health, University Of Manchester*

Background: When clinical prediction models (CPMs) are updated, they may encounter a feedback loop if the update incorporates data from individuals whose treatment was influenced by the model. This can lead to a prediction paradox, where variables initially predictive of poor outcomes become linked to treatment and better outcomes. The impact of this paradox on model performance over time, and how it varies across different clinical settings, is not well-documented. One proposed solution to this issue is to include treatment as a predictor in the models. Therefore, this study aims to explore how CPMs perform when they are updated with data influenced by the CPM itself and to determine whether incorporating treatment as a predictor can help counteract the prediction paradox.

Methods: A simulation study was carried out in which a logistic regression CPM was developed and underwent several implementation and update cycles using recalibration. Three parameters (outcome prevalence, impact of the treatment on patient outcomes and risk threshold for receiving treatment during implementation) were varied to create different scenarios. This process was repeated with a CPM that included treatment status as a predictor. Measures of predictive performance (discrimination and calibration) were calculated throughout cycles and averaged across the 200 bootstrap samples for each scenario.

Results: After being updated, the CPM underestimated risk and showed poor predictive accuracy for untreated patients. Performance was worse with increasing outcome prevalence, decreasing the risk threshold for treatment, and increasing the risk reduction with treatment. After the second update cycle, the model was well calibrated when validated after the treatment implementation strategy had been applied. Including treatment as a predictor variable improved calibration for the untreated but the model still underestimated risk on average.

Conclusion: The prediction paradox can cause risk to be underestimated for individuals resulting in missed opportunities for treatment. It is advisable to thoroughly evaluate the implications of this paradox when developing, implementing, updating, and validating CPMs, particularly in selecting appropriate populations for these activities. Adjusting for treatment is not a valid strategy for preventing the impact of the prediction paradox and more advanced modelling methods should be considered.

O30. CHARIOT: A prediction-under-intervention model for cardiovascular primary prevention

Matthew Sperrin[1], Bowen Jiang[1], Joyce Huang[1], Brian McMillan[1], Alexander Pate[1]
[1]*University Of Manchester*

Background: In current clinical prediction models in cardiovascular disease primary prevention, the emphasis is on estimating absolute risk of future outcomes, and proposing intervention where risks are high. However, this 'risk-based approach' does not directly consider the benefits of interventions (e.g., not providing risks for 'on' versus 'off' treatments), limiting ability to support decisions. New methods in prediction under intervention (counterfactual prediction) could address this.

Objectives: We introduce CHARIOT (Cardiovascular Health Assessment and Risk-based Intervention Optimisation Tool), a prediction under intervention model for cardiovascular disease prevention, that will allow assessment of both pharmaceutical and lifestyle interventions.

Methods: In an iterative approach, we are developing models using 20 million individuals in CPRD. First, a 'standard' clinical prediction model was developed as a comparator. Prototype 1 then accounts for treatment drop-in through adjustment to post-baseline hazard ratios, utilising trial- estimated causal effects. Prototype 2 addresses an issue regarding predictor variables which are impacted by the intervention (e.g., we should be explicit about whether measured blood pressure has been achieved through anti-hypertensive medication or not). Prototype 3 captures the effect of interventions via their effects on predictor variables, similar to mediation analyses, where effects of predictor variables are fixed to causal values.

Results: The 'standard' model generates risk in between risks 'on' and 'off' treatment under the prediction under intervention models. In Prototype 3, a penalty in predictive accuracy is necessary for the required causal interpretations.

Conclusions: This work has generated new methodological questions throughout. To equip models with causal interpretations requires strong assumptions which are often untestable. For example, Prototype 3 makes very strong assumptions regarding consistency. However, Prototype 3 is scalable and allows any intervention (including lifestyle) to be included, provided effects on predictor variables are understood. Future work will further allow for treatment interaction and heterogeneity, better triangulating between experimental and observational data.

A partner project is exploring how best to communicate these concepts to patients and a front-end interface is in development.

O31. Stronger penalties on treatment-covariate interactions improve treatment effect predictions and prevent potential treatment mistargeting

David Van Klaveren[1,3], Ewout Willem Steyerberg[2], David Michael Kent[3]
[1]Department of Public Health, Erasmus MC University Medical Center, [2]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, [3]Predictive Analytics and Comparative Effectiveness Center, Tufts Medical Center

Background: Counterfactual prediction models with treatment-covariate interactions can identify patient groups with differential treatment effects, but often result in exaggerated variation in treatment effect predictions. LASSO regression does not fully prevent this overestimation of treatment effect heterogeneity.

Purpose: To develop shrinkage techniques that reduce overestimation of treatment effect heterogeneity.

Methods: For "Interaction shrinkage", we first decomposed model predictions on the logit scale into a prognostic index and a deviation from the prognostic index due to treatment. We then used bootstrap validation to estimate shrinkage factors for these components. Alternatively, "Interaction penalization" uses penalized regression (LASSO) with a multiplicative penalty factor for interaction effects versus 1 for main effects. This factor is based on the sum of the interaction effects divided by the sum of the main effects in a preliminary model fit. We exploited a previously published simulation framework (samples of 3,600 patients; 25% binary outcomes; 12 binary covariates) – both without and with 6 true treatment-covariate interactions – to compare these approaches with a standard LASSO model, and with a Causal Forest. We measured calibration by the median difference between predicted and true treatment benefit in the first and fourth quarter of predicted treatment benefit (extreme-quarter-calibration; "EQC"; ideally 0). We measured discrimination by the median difference between the true treatment benefit in the fourth and first quarter of predicted treatment benefit (extreme-quarter-discrimination; "EQD", larger is better).

Results: WITHOUT true treatment-covariate interactions, the interaction shrinkage (median 0.36) and interaction penalty factor (median 2.32) were important. Compared to Causal Forest (EQC 5.7%; EQD 0.5%) and LASSO (EQC 2.2%; EQD 0.8%), predictive performance of Interaction shrinkage (EQC 2.0%; EQD 1.5%) and especially Interaction penalization (EQC 0.6%; EQD 2.5%) was better.
WITH true treatment-covariate interactions, the interaction shrinkage (median 0.60) and interaction penalty factor (median 1.62) were still considerable. Even WITH true treatment-covariate interactions, calibration of Interaction shrinkage (EQC 1.2%) and Interaction penalization (EQC 1.3%) was better compared to LASSO (EQC 1.9%) and much better compared to Causal Forest (EQC 3.8%; EQD 7.2%).

Conclusion: Data-driven techniques targeted at shrinkage of treatment-covariate interaction effects improve a model's ability to predict treatment effects.

O32. Effects of Using Natural Language Processing for Cohort Selection from Electronic Health Records on Subsequent Prognostic Prediction Model Performance

Isa Spiero, J.A.A. Damen[1,2], Dr. L. Hooft[1,2], Dr. K.G.M. Moons[1], A.M. Leeuwenberg[1]
[1]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, [2]Cochrane Netherlands, University Medical Center Utrecht, Utrecht University

Background: Electronic Health Records (EHRs) provide enormous amounts of data that could be utilized for research purposes, such as prognostic prediction modeling. Much valuable data in EHRs are present as unstructured data in the form of clinical notes. Extraction of these data can be automated by applying Natural Language Processing (NLP). However, erroneous NLP models can extract incorrect data which can in turn lead to bias if these data are used for subsequent prediction modeling.

Objectives: We aimed to evaluate the use of NLP for extraction of data for development of prognostic models. We specifically focused on the effect of erroneous NLP models in selection of patients for inclusion (i.e., cohort selection) from EHRs on the performance of the prediction model.

Methods: We retrieved EHR data from the semi-publicly available MIMIC-III database, containing data of over 40,000 ICU patients (Johnson et al., 2016). We trained NLP models of decreasing performance to extract 50 different disease states of patients from their clinical notes. ICD-codes present in the dataset were used as reference standard. Using these NLP models, patients were selected and thereby we created various potentially biased (i.e. falsely included and/or excluded patients) cohorts. These cohorts were used to develop a logistic regression prediction model, with the predictors from the Simplified Acute Physiology Score (SAPS) and quick Sepsis-related Organ Failure Assessment (qSOFA) to predict in-hospital mortality. The performance of the NLP models in selecting patients (expressed by the F1-score) was then related to the performance of the prediction model (expressed by the AUC, calibration intercept and slope, and model coefficients).

Results: Overall, NLP models with decreasing cohort selection performance also selected increasingly biased cohorts of patients, and consequently the prediction model coefficients varied between these selected cohorts. However, for most disease states, these biased cohorts did not affect the AUC or calibration when used in subsequent prediction model development (Figure).

Conclusions: The performance of an NLP model does affect its capability to correctly extract cohorts of eligible patients for prediction modeling and the resulting model coefficients. However, biased cohorts did not affect the overall performance of prediction models in our data.

O33. Identifying Priority Areas for Target Product Profile Development in Early Cancer Diagnostics

Jac Dinnes[1], Mary Jordan[2], Pranshu Mundada[2], Jessica Lloyd[3], Emma Jobson[3], Sowmiya Moorthie[3], Samantha Harrison[3], **Bethany Shinkins**[2]
[1]University of Birmingham, [2]University of Warwick, [3]Cancer Research UK

Introduction: The NHS aims to diagnose 75% of all cancers at stage I or II by 2028. To achieve this, stimulating research and innovation in early cancer diagnostics will be essential. Target Product Profiles (TPPs), which outline the essential characteristics of a new health technology may provide a useful mechanism. However, TPP development efforts need to be focused on areas of high clinical priority. We aimed to identify priority areas of unmet need for novel early cancer diagnostics to guide future TPP development activity.

Methods: We used an iterative process to develop a set of prioritisation criteria. A set of burden of disease metrics provided by Cancer Research UK for 22 cancers was updated. Four trade-off exercises were developed and presented at a conference workshop to identify additional prioritisation criteria. The updated criteria were then split into four categories: 1) Incidence, Mortality, and Survival, 2) The Case for Earlier Diagnosis, 3) Current Diagnostic Pathways, and 4) Horizon Scanning. A slide deck and exercise were presented to an expert stakeholder group. Participants were prompted to categorise each cancer as higher, medium or lower priority areas for test innovation, with no limits on number per category. Responses were scored from 3 (higher) to 1 (lower priority), with an average score calculated for each cancer. Priority groups were formed based on score profile and clustering.

Results: Twenty expert stakeholders completed the prioritisation exercise. Consensus was achieved on the lower priority group (larynx, cervix, testis and Hodgkin lymphoma). In the medium priority group, six cancers were lower-medium priority (non-Hodkin lymphoma, brain, breast, melanoma, leukaemia, and myeloma) and two were higher-medium (colorectal and liver) due to the averaging effect from a wider split of responses across all priority levels. Six cancers (upper GI, pancreatic, urological, gynaecological, prostate and lung) formed the higher priority group, although extent of agreement varied by cancer. The driving factors for prioritisation differed for each cancer and will be discussed.

Conclusions: This prioritisation exercise provides valuable insights into the areas where new early diagnostic tests for cancer could have the most significant impact on health outcomes and the health care system.

O34. Developing diagnostic target product profiles for managing infections and exacerbations in cystic fibrosis: a sequential mixed-methods design.

Nicola Howe[1], Kile Green[1], Constance Takawira[3], Lorna Allen[4], Neill Gingles[3], Paula Sommer[4], Raasti Naseem[2], Rachel Dakin[2], Rebecca Holmes[2]
[1]NIHR HealthTech Research Centre (HRC) in Diagnostic and Technology, [2]LifeArc, [3]Medicines Discovery Catapult, [4]Cystic Fibrosis Trust

Background: Optimising and preserving lung function is key to maintaining good clinical outcomes for people with Cystic Fibrosis (pwCF). Timely and appropriate detection, identification, management, and informed treatment of complications such as exacerbations and infections is central to achieving this aim, preventing disease progression, and minimising further complications. Challenges regarding symptom variability, non-symptomatic infections, poor sample availability, and culturing delays present an opportunity for the Cystic Fibrosis (CF) community and diagnostic developers to work together to support the development of patient-centred diagnostics for exacerbation and infection in CF which address unmet diagnostic needs.

Objectives: To develop patient-focused diagnostic target product profiles (TPP) addressing unmet needs in managing infections and pulmonary exacerbations in cystic fibrosis. A diagnostic TPP outlines the necessary characteristics, qualities, and clinical utility that diagnostic tests should demonstrate in order to address an unmet clinical need.

Methods: This project utilised an exploratory sequential mixed-methods approach in three stages, guided by an expert advisory group of stakeholders including methodologists, patients, developers and clinical staff.

Stage 1- a landscape analysis to identify and prioritise unmet diagnostic needs via multiple focus groups of clinical experts and pwCF, plus a scoping review of the diagnostic space and available diagnostic tests.
Stage 2- drafting TPPs using existing literature and regulatory documentation, evidence from the focus groups, one-to-one interviews with stakeholders, and web-based surveys, to define 'minimal' and 'optimal' characteristics for each TPP.
Stage 3- refined and validated TPP content through additional interviews, a two-round modified Delphi exercise, and a virtual symposium.

Results: A comprehensive document (available freely via the CF AMR Syndicate website) aimed to guide diagnostic developers in research for pwCF was created and released freely. The document includes 'high-level' TPPs covering diagnostics for managing acute pulmonary exacerbations, rapid pathogen identification, and antimicrobial susceptibility tests, as well as a more detailed TPP defining in vitro diagnostic tests for rapid detection of non-tuberculous mycobacteria (NTM) pulmonary infections.

Conclusions: The project consulted over 150 individuals and experts in CF management, infection, and diagnostic development. The TPP guidance document supports research and development of patient-focused diagnostics for the benefit of pwCF and healthcare systems.

O35. Lost in Translation: The Current and Future Regulatory Landscape as an Often-Overlooked Hurdle for Impact in Clinical Prediction Models

Benjamin Perry[1]
[1]*Institute For Mental Health, University Of Birmingham*

There is a critical health and economic need for improved mechanisms of primary disease prevention in healthcare. Clinical Prediction Models (CPMs), when implemented, are a demonstrable means to facilitate primary prevention, and are increasingly prioritised by academic researchers and public grant funding bodies in the UK and beyond. Yet, there is an alarming lack of translation of published CPMs into routine practice, and most published CPMs contribute nothing more than research waste. While steps to improve methodological rigour are now common-place, steps to help researchers in academic settings clear regulatory hurdles, which are a pre-requisite toward implementation and impact, are lacking.

This talk will focus on the regulatory landscape as it has existed, and how it is shifting both in the UK and internationally. The talk will consider the strengths and limitations of current and proposed future frameworks, considering the particular hurdles faced by CPMs developed within academia. CPM case studies of success and failure will be presented to highlight the particular barriers faced by CPMs developed within academia.

Three inter-dependent big asks are proposed for academic institutions and funders that could help to address this problem, maximising the chances that the large and ever-increasing investment by public funding bodies toward CPM research in academia translates into meaningful patient benefit.

O36. Assessment of Prediction Models in Europe: Gaps in Evidence Requirements

Tuba Saygin Avsar[1], Alina Solomon, Miia Kivipelto, Francesca Mangialasche, Niranjan Bose, Dalia Dawoud, Bethany Shinkins
[1]NICE

Introduction: Clinical prediction models combine various types of information e.g. patient characteristics, symptoms, biomarker and imaging results, to support clinical decision making and improve patient's health outcomes. As these models gain popularity, clear HTA processes are essential to ensure those which improve outcomes reach clinical practice. Here, we identify HTA guidance and processes for evaluating clinical prediction models across Europe.

Methods: HTA methods manuals published by HTA agencies in countries within the European Economic Area were identified. Those focused on screening, diagnosis and companion diagnostics were reviewed to identify any available guidance on the HTA of clinical prediction models. Data were analysed using the EUnetHTA Core model framework. A series of online workshops were organised with HTA experts from 14 different organisations across 10 European countries to discuss their experience of evaluating clinical prediction models to date, the challenges faced, and to identify existing adoption pathways for novel clinical prediction models.

Results: The review found 29 potentially relevant methods guidance documents. EunetHTA, HAS, IQWIG, NICE, SBU, and ZiN offered the most comprehensive and relevant guidance on the assessment of diagnostics more generally, with only 3 (IQWIG, NICE and SBU) providing some guidance specific to clinical prediction models. Workshop discussions supported these findings, as HTA agency representatives shared limited experience with the assessment of clinical prediction models, and a lack of clear adoption pathways. Decision-critical issues with insufficient guidance were identified as follows: Validation and applicability assessment, context of use, link to patient outcomes, and handling uncertainty.

Conclusions: This study provides an overview of the current landscape regarding HTA agencies' guidance and processes for the HTA of clinical prediction models across Europe. Experience in the HTA of clinical prediction models was very limited. The findings highlight the need for further methodological guidance on the assessment of clinical prediction models and clearer adoption pathways.

Poster Presentation Abstracts

P1. Stop before you start: a checklist for those thinking about developing a clinical prediction model

Lucinda Archer[1,2], Rebecca Whittle[1,2], Kym Snell[1,2], Paula Dhiman[3], Gary Collins[3], Richard Riley[1,2], Joie Ensor[1,2]
[1]Department of Applied Health Sciences, University of Birmingham, [2]National Institute for Health and Care Research (NIHR) Birmingham Biomedical Research Centre, [3]Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford

Background: Clinical prediction models are currently being published at an alarming rate, but most are of poor quality and have no impact on clinical practice. A major reason for this is that models are often developed with little prior consideration of clinical need, study design, and potential requirements for implementation. This ultimately hampers their predictive performance, uptake, and acceptability.

Methods: To improve research standards and building on previous work, we describe a new checklist of key issues to consider prior to developing a clinical prediction model. This checklist was formed through consultation and consensus among experts in the prediction modelling field, both methodological and clinical. We outline the potential impact of neglecting each item on the predictive performance, clinical utility, and acceptability of the developed model.

Results: The checklist covers key issues researchers should consider when planning to develop a clinical prediction model and is applicable irrespective of clinical area or modelling approach.
Items surrounding study design include: recruitment of and consultation with a relevant Patient and Public Involvement and Engagement (PPIE) group; full consideration of the research question under investigation (e.g., using PICOTS: Population, Implementation, Competing models, Outcomes, Timing, Setting) with input from key stakeholders; and identification of the necessary sample size to develop a stable and generalisable model.
Regarding consideration of model implementation, items cover: ensuring model outcomes and potential predictors are both measurable and acceptable; understanding why any existing models in the field are not currently being used, and what improvements are needed; and assessing potential barriers to implementation in clinical practice, both on a population-level and in key subgroups.

Conclusion: Prediction model development should be preceded by crucial preparatory work to help ensure that the resulting models are statistically robust and clinically relevant. The checklist should be considered at the pre-protocol stage of a research project and can help guide researchers on whether they should be developing a prediction model or not.

Overall, the checklist aims to ensure investigators achieve maximum relevancy and utility of their clinical prediction models, while reducing the risk of their work becoming research waste.

P2. Improving the reference standard in diagnostic accuracy studies: Evaluating a latent class model against a panel of expert clinicians

Joie Ensor[1, 2], Kym Snell[1, 2]
[1]Department of Applied Health Sciences, School of Health Sciences, College of Medicine and Health, University of Birmingham, [2]National Institute for Health and Care Research (NIHR) Birmingham Biomedical Research Centre

Background: Flexible calibration curves are essential tools for assessing clinical prediction model (CPM) performance and are recommended in the recent TRIPOD+AI guidelines. Despite their recommendation, no guidance is given on how best to construct calibration curves and there is limited evidence comparing different smoothing approaches for their construction. While methods such as a locally weighted least squares regression smoother (loess) are commonly used in available software, the relative merits of alternative smoothing methods in detecting different types of miscalibration and computational efficiency remain unexplored, particularly in small/large datasets and in rare diseases.

Objective: To systematically compare smoothing methods for calibration curves, evaluating their: (1) ability to detect different types of miscalibration, (2) computational efficiency, (3) stability in regions with varying data density, and (4) sensitivity to tuning parameters.

Methods: We evaluated multiple smoothing approaches including loess, restricted cubic splines and kernel-based smoothing. Using both simulated and real clinical data, we assessed each method's performance in detecting varying degrees of systematic miscalibration and overfitting. We compared smoother performance in terms of efficiency and stability and investigated adaptive weighting strategies to optimise computational efficiency in regions with varying data density.

Results: Our findings revealed distinct advantages for different smoothing approaches across various prediction scenarios. Traditional methods like loess demonstrated robust performance across standard settings, while alternative approaches showed superior computational efficiency in larger datasets. We identified optimal approaches for challenging scenarios including small validation studies and rare disease settings, where data density varies substantially across the risk spectrum. We observed varying sensitivity to parameter choices (e.g., span in loess, knot placement in splines) affecting the balance between flexibility and stability.

Conclusions: We provide practical guidance for choosing appropriate smoothing methods in calibration assessment, considering both statistical performance and computational efficiency. These results have important implications for the routine assessment of prediction model calibration, particularly in extremes of validation study size and in rare diseases.

P3. Calibration plots for clinical prediction models predicting time-to-event outcomes, a focus on the role of censoring

Alexander Pate[1], Matthew Sperrin[1]
[1]*Division of Imaging, Informatics and Data Science, University Of Manchester*

Background: There are several methods for estimating calibration plots for a model predicting a time-to-event outcome at a specific time point t, including proportional hazards regression, inverse probability of censoring weights (IPCW), and pseudo-values. The assumptions each of these methods make, and how they perform under different censoring mechanisms has not been compared in detail.

Objectives: To introduce these methods to a broader audience, showcase each of the methods in reproducible examples, and identify when each of these methods give a biased assessment of calibration.

Methods: Data was simulated for individuals with three predictors (x1, x2 and x3). Time-to-event outcomes were simulated according to exponential survival distributions with three different hazards: single linear (SL), multiple linear (ML) and multiple non-linear (MNL). A time until censoring was also simulated according to SL, ML and MNL structures, and an independent censoring mechanism. Development and validation datasets of size 500,000 were simulated. A SL, ML and flexible model were fitted to the development data, and calibration curves were estimated using the proportional hazards regression, IPCW and pseudo-value approaches. We compared the estimated calibration curves with true calibration curves calculated from the data generating mechanism.

Results: Under independent censoring, all the methods give unbiased assessments of calibration. If a variable (or non-linear transformation of a variable) is omitted from a prediction model, and this variable is predictive of both the outcome and censoring mechanism in the validation dataset, calibration curves estimated using the proportional hazards or pseudo-value (grouped by predicted risk) approaches will be biased. The ability of the IPCW and pseudo-value (grouped by IPCWs) approaches to assess calibration were not dependent on which variables were included in the prediction model.

Conclusions: If a developed model has predictors omitted (e.g. due to sample size restrictions, or predictors not being available at model implementation), which are predictive of both the outcome and censoring mechanism, extra care is needed to gain an unbiased assessment of calibration. For some approaches to calibration, the ability to assess calibration of the clinical prediction model is tied to the specification of the clinical prediction model itself.

P4. Validating recurrent event clinical prediction models using epilepsy as an example.

Alexandra Hunt[1], Thomas Spain[1], Laura Bonnett[1], Hein Putter[2], Anthony Marson[1]
[1]University Of Liverpool, [2]Leiden University Medical Centre

Background / Introduction: Prediction models for recurrent medical episodes, such as seizures, are increasingly being developed but often lack thorough validation. External, and internal validation, is crucial for assessing model performance, in wider populations, through calibration and discrimination. A systematic review of 301 recurrent event models found only 24.9% were internally validated, and just 1.0% underwent external validation, highlighting methodological challenges in this area.

Recurrent event models require specialised approaches for validation. Calibration can utilise predicted and observed event counts rather than probabilities, but discrimination methods must account for the varying nature of recurrent events, which is often overlooked.

We developed a discrimination method for recurrent event models and applied it, alongside modified calibration techniques, to models predicting future seizure counts in people with epilepsy.

Method: Using SANAD data, we built Andersen-Gill (AG) and Prentice-Williams-Peterson (PWP) Cox models and validated them externally using SANAD-II data, which included 1,510 patients. Calibration was assessed via event count-based plots and observed-to-expected ratios. Discrimination was evaluated using a concordance measure comparing predicted and observed event counts.

Results: Results show better calibration for the PWP model compared to the AG model. Internal and external discrimination results for the PWP model are excellent, with c-statistics reaching 0.94 and 0.85, respectively. Full results, including differences by model and epilepsy type, will be presented.

Conclusion: Our methods enhance current methods of validating recurrent event models, a critical step for improving care in conditions characterised by repeated episodes, such as epilepsy. Future work includes integrating these methods into R software to support researchers.

P5. Developing prognostic models in rare diseases: A systematic review of sample size and methodological approaches in recent studies

Laura Kirton[1], Richard Riley[2,3], Piers Gaunt[1], Kym Snell[2,3]
[1]Cancer Research UK Clinical Trials Unit, College of Medicine and Health, University of Birmingham, [2]Department of Applied Health Sciences, School of Health Sciences, College of Medicine and Health, University of Birmingham, [3]National Institute for Health and Care Research (NIHR) Birmingham Biomedical Research Centre

Background: Current sample size guidance for developing prognostic models in a specific target population aims to minimise overfitting, minimise optimism in R2 and target precise estimation of the overall risk. In the setting of rare diseases, reaching this minimum sample size can be unachievable; yet there is still appetite from patient groups and clinicians for individualised risk predictions, which raises methodological challenges for model development, stability and clinical utility.

Objectives: To undertake a systematic review of recent studies aiming to develop a prognostic model in a rare disease setting. Specifically, to investigate (lack of) adherence to the existing sample size guidance, summarise if and how sample sizes were determined, highlight methodological approaches being used to address sparse data during model development, and identify whether uncertainty and instability of predictions are examined.

Methods: The review focuses on the clinical exemplar of sarcomas, as all sarcoma histological subtypes are classed as rare cancers. Through a search of MEDLINE and Embase, publications since January 2019 to August 2024 of prognostic models developed in sarcomas were identified. Key data were extracted from each publication, including sample size information, methods and justification for the model development approach, and whether instability or uncertainty of model predictions and clinical utility were considered. Additionally, where authors employed statistical methods or approaches to address the limited sample size issue, a qualitative description was extracted.

Results: Preliminary findings suggest many publications do not consider or address sample size at all, and most do not adhere to existing sample size guidance. To summarise this problem, the size of development datasets being used and model development approaches will be summarised. Statistical methods being utilised to address issues of small sample sizes will be presented and critiqued. Additionally, examples will be given to showcase (concerns of) instability and imprecision.

Conclusions: The findings will provide an important overview of current prognostic model development studies in rare diseases, highlighting concerns that sample size limitations are often ignored. The prevalence of rare diseases results in limited available sample sizes for prognostic modelling studies and therefore there is a need for alternative statistical methodology in these scenarios.

P6. Updating methods for AI-based clinical prediction models: a scoping review

Lotta M. Meijerink[1], Zoë S. Dunias[1], Artuur M. Leeuwenberg[1], Anne A.H. de Hond[1], David A. Jenkins[2], Glen P. Martin[2], Matthew Sperrin[2], Niels Peek[3], René Spijker[1], Lotty Hooft[1], Karel G. M. Moons[1], Maarten van Smeden[1], Ewoud Schuit[1]

[1]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, [2]Division of Informatics, Imaging and Data Sciences, University of Manchester, [3] The Healthcare Improvement Studies Institute, Department of Public Health and Primary Care, University of Cambridge

Background: Prediction models may not always generalize well across different contexts, e.g. time periods, hospitals or medical domains. Adjusting an existing prediction model to new situations ('updating') based on new data is considered an efficient approach, but an overview of updating methods for AI-based clinical prediction models is lacking.

Objectives: To summarize methods for updating AI-based clinical prediction models.

Methods: We comprehensively searched Scopus and EMBASE up to August 2022 for articles that addressed developments, descriptions, or evaluations of prediction model updating methods. We focused on articles in the medical domain involving AI-based prediction models, excluding regression-based updating methods. We categorized and described the identified methods used to update the AI-based prediction model as well as their use cases.

Results: We included 78 articles. Most articles discussed updating for neural network methods (93.6%) with medical images as input data (65.4%). In many articles (51.3%) existing, pre-trained models for broad tasks were updated to perform specialized clinical tasks. Other common reasons for model updating were to address changes in the data over time and cross-center differences; however, more unique use cases were also identified, such as updating a model from a broad population to a specific individual.

We categorized the identified model updating methods into four categories: neural network-specific methods (described in 92.3% of the articles), model-agnostic methods (9.0%), ensemble-specific methods (2.5%), and other (1.3%). Variations of neural network-specific methods are further categorized based on: (1) the part of the original neural network that is kept, (2) whether and how the original neural network is extended with new parameters, and (3) to what extent the original neural network parameters are adjusted to the new data. The most frequently occurring method (n=30) involved selecting the first layer(s) of an existing neural network, appending new, randomly initialized layers, and optimizing the entire neural network.

Conclusions: We identified many ways to update AI-based prediction models, across various use cases. Updating methods for AI-based prediction models other than neural networks (e.g., random forest) appear to be underexplored in clinical prediction research.

P7. Teaching critical appraisal of clinical prediction model studies to Health Data Science students using the Prediction model Risk Of Bias ASsessment Tool (PROBAST) as a teaching tool

Jamie Sergeant[1]
[1]*Centre for Biostatistics, University Of Manchester*

Background: It is well documented that deficiencies in methodology and reporting are common in studies which develop and/or validate clinical prediction models (CPMs). It would be beneficial for trainee health data scientists to be able to recognise and understand these deficiencies, enabling them to assess the value of studies they will encounter in their careers, and informing the design and reporting of their own studies. However, critical appraisal of CPM studies is a topic that may not readily appeal to students perhaps more motivated to study advanced analytical methods.

Objective: To teach critical appraisal of CPMs to postgraduate Health Data Science students, using 1) the application of machine learning in CPMs as a hook to attract students, and 2) the Prediction model Risk Of Bias ASsessment Tool (PROBAST) as a teaching tool.

Methods: Teaching was delivered in 2022-23 and 2023-2024 as part of an optional 15-credit module on advanced statistical topics. The module aimed to foster statistical thinking, enable engagement with published research, and promote informed discussion and debate. An active learning approach was used, whereby students engaged with research through directed reading and enquiry as independent study. In-person classes were used for the reinforcement of core content, practical exercises and group presentations. Experiential and research-led learning saw students undertake their own PROBAST risk of bias assessments on published studies, first as a formative exercise, then as a summative assessment, delivered as an exam in 2022-23 and as an individual coursework assignment in 2023-24.

Results: The number of students grew from 22 in the first year to 100 in the second year. It was a challenge to enable and empower all students, including the high proportion of overseas students, to engage in whole-class discussions. Online interactive whiteboards were a valuable tool in addressing this challenge. Almost all students were able to competently perform a PROBAST risk of bias assessment and adequately justify their decisions. The use of generative AI was an issue in coursework assessment.

Conclusion: Health Data Science students developed their statistical thinking skills through the authentic application of PROBAST to research literature.

P8. Evaluating the discrimination of prediction models for recurrent medical events

Thomas Spain[1], Alexandra Hunt[1], Hein Putter[2], Victoria Watson[1], Laura Bonnett[1]
[1]Department of Health Data Science, University Of Liverpool, [2]Leiden University Medical Center, University of Leiden

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

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

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

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

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

P9. Extending sample size calculations for evaluating clinical prediction models that use a threshold for classification

Rebecca Whittle[1,2], Joie Ensor[1,2], Lucinda Archer[1,2], Gary S. Collins[3], Paula Dhiman[3], Alastair Denniston[2], Joseph Alderman[4], Amardeep Legha[1,2], Maarten van Smeden[5], Karel G. Moons[5], Jean-Baptiste Cazier[6], Richard D. Riley[1,2], Kym I.E. Snell[1,2]
[1]Department of Applied Health Sciences, School of Health Sciences, College of Medicine and Health, University of Birmingham, [2]National Institute for Health and Care Research (NIHR) Birmingham Biomedical Research Centre, [3]Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, [4]Department of Inflammation and Ageing, School of Infection, Inflammation and Immunology, College of Medicine and Health, University of Birmingham, [5]Julius Center for Health Sciences and Primary Care, University Medical Centre Utrecht, Utrecht University, [6]Francis Crick Institute

Background: When evaluating (validating) the performance of a model for risk prediction, the sample size needs to be large enough to precisely estimate the performance measures of interest. Current sample size guidance is based on precisely estimating calibration, discrimination, and net benefit, which should be the first stage of calculating the minimum sample size required. However, when a clinically important probability threshold is used for classification, other performance measures are also relevant.

Objectives: We extend previously published sample size guidance to target precise estimation of threshold-based performance measures, to inform the sample size required for studies evaluating a prediction model with a binary outcome.

Methods: We derive closed-form solutions to estimate the minimum sample size required to target sufficiently precise estimates of accuracy, specificity, sensitivity, PPV and NPV, and an iterative method to estimate the sample size required to target a sufficiently precise estimate of the F1-score. The calculations require the user to pre-specify target standard errors (confidence interval widths) and provide either (i) the assumed true value of each performance measure or (ii) a large synthetic (simulated) dataset that approximates the assumed distribution of the model's predicted risk in the target population, which allows the true values to be calculated at any relevant probability threshold.

Results: We describe how the sample size formulae were derived and showcase our corresponding R, Stata and Python software package (pmvalsampsize). We provide a worked example of an external validation study planning to evaluate a published model for predicting in-hospital clinical deterioration. In this example, the minimum sample size required was lower than that required to precisely estimate the calibration slope, and we expect this would often be the case. Extension to time-to-event outcomes is also considered.

Conclusions: Our formulae, along with pmvalsampsize, enables researchers to calculate the minimum sample size needed to precisely estimate threshold-based performance measures in an external evaluation study. When classification is relevant based on risk thresholds, these additional criteria should be used alongside existing criteria to precisely estimate calibration, discrimination, and net benefit.

P10. Developing a prediction model for prostate cancer patients on MRI-led Active surveillance for progression to treatment

Busola Adebusoye[1], Cameron Englman[2], Associate Professor Giganti Francesco[2], Caroline Moore[2], Sue Mallett[1]
[1] Centre for Medical Imaging, University College London, [2]Division of Surgery & Interventional Science, University College London

Background: Patients on active surveillance for localised prostate cancer differ in their likelihood of needing treatment, and intensity of follow up could be tailored so that those with lowest likelihood have fewer investigations.

Objective: To develop a prediction model for prostate cancer (PCa) patients on active surveillance (AS), to indicate a likelihood of needing treatment by 5 years, according to baseline characteristics.
Materials & Methods: PCa patients from a consecutive clinical cohort of men on MRI-led active surveillance (University College London Hospital, January 2005 date to July 2023 date), were included. Each had a matched MRI and biopsy showing Gleason 3 + 3 or Gleason 3 + 4 prostate cancer, and at least one follow-up MRI.

A multivariable flexible parametric model was developed with 1047 men. Based on an expert consensus process, a small number of key predictors were included: Gleason score, MRI visibility (Likert≥4), PSA density and maximum cancer core length (MCCL). Model calibration and discrimination were completed alongside internal validation (100 bootstraps). Results were presented as linear predictors and 5-year survival probability and calibration plots at 5 years. Example predicted survival curves for typical individual patients were also reported.

Results: 1047 PCa patients (median age 63, IQR (57,68) years) with the outcome progression to treatment (n=339) were included in the model. The model discrimination was reported as C-statistic of 0.75 (95%CI 0.72, 0.77) and calibration showed good performance at both high and low risk. A simple chart was constructed to provide the 5-year survival probability of the patients remaining on active surveillance based on their four baseline clinical parameters. Example patient predictions of treatment-free survival are presented.

Conclusion: This simple and transparent model will be useful for patient communication, decision making on monitoring tests, frequency of monitoring and to guide further treatment options.

P11. Predicting the risk of and days without albuminuria in patients with diabetes mellitus: a development and validation study

Roemer Jonah Janse[1], Chava L. Ramspek[1], Marretje W. Oomen[1], Friedo W. Dekker[1], Juan-Jesus Carrero[2], van Diepen[1]
[1]Leiden University Medical Center, [2]Karolinska Institutet

Background: Patients with type 2 diabetes mellitus (T2DM) are at high risk of kidney disease, which may be ameliorated by early detection of albuminuria. Clinical prediction models (CPMs) can tailor guideline-indicated screening to the individual patient, but no adequate CPMs are currently available.

Objectives: We aimed to predict 3-year albuminuria risk, albuminuria-free time, and progression through albuminuria stages over 3 years at the moment of a first normoalbuminuric test.

Methods: We used data from the Stockholm Creatinine Measurements cohort. We selected patients with T2DM, a normoalbuminuric test, and no prior albuminuria between 2007-2021 from all Stockholm residents. Subsequently, we created a development (2007-2013) and temporal validation (2014-2021) cohort. Predictors were selected based on clinical expertise, literature, and previous CPMs. Albuminuria was defined as urine albumin-creatinine ratio (uACR) ≥30mg/g. We predicted an individual's risk of albuminuria using a Fine-Gray CPM taking into account the competing risk of death and showing risk progression over time. Additionally, we predicted the albuminuria-free time using an accelerated failure time (AFT) CPM. Lastly, we predicted the risk of transitioning to microalbuminuria (uACR ≥30mg/g & <300mg/g), to macroalbuminuria (uACR ≥300mg/g), and to death using a multistate CPM. Model discrimination and calibration were assessed in the development and temporal validation cohort.

Results: The development cohort contained 38,649 individuals with 6,904 events and the validation cohort contained 45,009 individuals with 6,499 events. The Fine-Gray CPM had adequate discrimination internally (C-statistic, 95%CI; 0.64, 0.64-0.65) and temporally (0.66, 0.66-0.67). Calibration was good. The Fine-Gray model was also able to accurately show individual albuminuria risk progression over 3 years. The AFT CPM had adequate discrimination (internally: 0.63, 0.62-0.63; temporally: 0.65, 0.64-0.65), but poor calibration. The multistate CPM allowed individual predictions for each state over 3 years. An example of an individual's predictions is shown in Figure 1.

Discussion: Predicting albuminuria in T2DM patients allows tailoring albuminuria screening to the individual. We developed multiple CPMs that accurately provide the probability of developing different stages of albuminuria over time. These models can serve to improve albuminuria screening and ameliorate the risk of kidney damage in patients with T2DM.

P12. Enhancing Alzheimer's Disease Clinical Trial Efficiency: Leveraging Prognostic Scores

Harry Parr[1], Doug Thompson[1], Jeffrey Lin[2], Dave Inman[1], Aris Perperoglou[1]
[1]GSK

Background: Alzheimer's Disease (AD) presents substantial heterogeneity in clinical presentation and progression of its disease course, this poses challenges in clinical trial efficiency and analysis. The advent of statistical and machine learning (ML) and extensive external data provides new opportunities to tackle these challenges by better understanding the drivers of patient heterogeneity.

Objectives: This study aims to demonstrate a novel methodology employing prognostic scoring adjustments to improve statistical efficiency and power for detecting treatment effects in AD clinical trials. By integrating these techniques into a repeated-measures framework, trial outcomes can be enhanced significantly.

Methods: We developed an ensemble of ML models trained on external data to generate prognostic scores (PS) for patients, to predict their standard-of-care outcomes. These PS were incorporated as 'super-covariates' within a Mixed Model for Repeated Measures (MMRM) analysis. Extensive simulations were conducted to quantify power improvements and assess robustness across different scenarios, including nonlinear relationships, heterogeneous treatment effects, and population shifts.

Results: The application of PS within the MMRM framework showcased significant power improvements, often exceeding several percentage points beyond the adjusted analysis, without compromising bias nor type I error. The methodology improved precision in treatment effect estimates, underscoring its potential in reducing sample sizes, lowering trial costs, and minimising patient burden.

Conclusions: Integrating ML techniques on historical real-world data to complement contemporaneous clinical trials by producing prognostic scoring in AD trials addresses heterogeneity and enhances trial efficiency. This approach significantly improves treatment effect precision and boosts power, facilitating the accelerated development of effective treatments. Incorporating these methods into the statistical analysis plan should become standard practice in AD trials to maximise trial success and efficiency.

P13. Evaluate design considerations when modelling time as continuous versus categorical in the presence of longitudinal Alzheimer's Disease clinical trial profiles.

Ashwini Venkatasubramaniam[1], Aris Perperoglou[1], Dave Lunn[1], Dave Inman[1], Katie Thorn[1], Doug Thompson[1]
[1]*GlaxoSmithKline*

Background: The Mixed Model for Repeated Measures (MMRM) is used ubiquitously in the analysis of longitudinal clinical trials across many indications in medicine research. There has been interest in using flexible natural cubic splines applied in Alzheimer's disease (AD) trials. A range of methods might be considered though, including regression splines, penalised splines, and fractional polynomials.

Objectives: This simulation study seeks to encourage greater embedding of continuous time models in clinical trials by describing the performance of such time models under multiple scenarios and highlighting their ability to leverage an interpretation of biological evolution. This evaluation offers an opportunity to bridge an identified gap in the literature.

Methods: We developed a comprehensive simulation study to compare the MMRM against alternative continuous time models, including: (i) natural cubic splines; (ii) penalised splines; or (iii) fractional polynomials as well as an "oracle" model as a benchmark. We simulated numerous scenarios for the true data generating mechanism according to realistic patterns of AD progression in clinical trials. We considered alternative study designs, ranging the number of visits and the sample size.

Results: We found that in the balance of type 1 error control and improvements in power, a natural cubic spline with one interior knot performed the best, achieving incremental gains in power ranging from 80-90% versus an MMRM that achieved 80% only. The fractional polynomial and the penalised spline, both of which harness some data-driven selection steps, suffered from some under-coverage induced via bias in the estimation of the model based standard error. In addition, the fractional polynomial suffered the most from bias in its estimates.

Conclusions: We recommend that a natural cubic spline offers strong potential in longitudinal continuous time models, though in the scenarios we explored there was an economical balance toward spending fewer knots. Alternative methods like the fractional polynomial and the penalised spline may be less suitable, particularly in application to superiority AD trials.

P14. Development and validation of a postpartum cardiovascular disease risk prediction model in women incorporating reproductive and pregnancy-related predictors.

Steven Wambua[1], Francesca L Crowe[1], Shakila Thangaratinam[1], Dermot O'Reilly[2], Colin McCowan[3], Sinead Brophy[4], Christopher Yau[5], Krishnarajah Nirantharakumar[1], Richard D Riley[1], Kym I E Snell[1]

[1]University Of Birmingham, [2]Queen's University Belfast, [3]University of St Andrews, [4]Swansea University, [5]University of Oxford

Background: Although recent evidence shows several pregnancy-related factors are associated with increased risk of cardiovascular disease (CVD), these factors are not usually included in current CVD risk prediction models.

Objectives: To determine whether adding pregnancy factors to a prediction model with established risk factors for CVD improves 10-year risk prediction of CVD in postpartum women using the QRISK®-3 risk equation as a benchmark model.

Methods: We used a population-based retrospective cohort of women aged 15 to 49 who have been pregnant from the Clinical Practice Research Datalink (CPRD) primary care database. We used established risk factors for CVD in the general population from QRISK®-3 such as age, ethnicity, deprivation, diabetes mellitus and some medication use. We then identified additional risk factors specific to women who have been pregnant such as gestational diabetes and hypertensive disorders of pregnancy from literature and from discussions with clinicians and patient research partners. We evaluated the performance of QRISK®-3 and further updated the risk prediction model for use specifically in postpartum women. First, we updated the baseline hazard, then updated the predictor effect estimates of established predictors, and finally considered additional pregnancy-related factors. Models were developed using Cox-proportional hazards regression for the outcome of CVD within 10 years. Models were evaluated and compared using measures of overall model fit, calibration, discrimination and clinical utility.

Results: Among 567,667 women who had been pregnant between 15-49 years of age, 2,175 (0.38%) experienced a CVD event within 10 years. Adding pregnancy factors to those from QRISK®-3 led to marginal improvements in model performance (QRISK®-3 C-statistic: 0.703 (95% CI 0.688 to 0.718), New model with traditional risk factors C-statistic: 0.716 (95% CI 0.701 to 0.731), New model with additional pregnancy factors C-statistic: 0.725 (95% CI 0.710 to 0.740), the clinical utility of models with additional pregnancy factors was better.

Conclusions: The updated risk prediction models resulted in marginal improvement in discrimination and calibration compared to QRISK®-3 in postpartum women. Although the overall predictive performance and calibration of the updated models was similar, the model with additional factors resulted in better clinical utility in women who had been pregnant.

P15. Assessing Adherence to TRIPOD+AI Guidelines in Machine Learning Models for Predicting SGA and FGR: A Systematic Review

Giulia Zamagni[1,2], Giulia Barbati[1]
[1]University of Trieste, [2]Epidemiology and Public Health Research Unit, Institute For Maternal And Child Health IRCCS Burlo Garofolo

Background: Fetal Growth Restriction (FGR) and Small for Gestational Age (SGA) are critical contributors to perinatal morbidity and mortality. Despite a 2016 consensus defining FGR, SGA remains widely used as a proxy, posing significant diagnostic challenges. The accurate prediction of FGR/SGA is crucial for timely interventions, but complex interactions can limit traditional statistical methods. Machine Learning (ML) offers promising solutions but requires methodological rigor. The TRIPOD+AI statement1 provides guidelines to improve transparency and generalizability of ML prediction models, and can be used to critically evaluate the existing scientific literature retrospectively.

Objectives: This review aims to evaluate the methodological rigor and adherence to standardized definitions and guidelines in studies employing ML models to predict FGR/SGA.

Methods: A systematic search was conducted in MEDLINE and Scopus following PRISMA 2020 guidelines. Studies were included if reporting ML models for FGR/SGA, with metrics like AUROC or accuracy. Minimum sample sizes in each study were determined using the approach of Riley2 for prediction models. As ML inherently demands more data than traditional statistical techniques, this established a "foundational minimum" for evaluating sample size adequacy. Adherence to TRIPOD+AI statement was assessed for each study using a 4-point Likert scale.

Results: Out of 272 identified records, 31 studies were included. A significant variability in outcome definitions was observed, with only 33.3% of studies adhering to the consensus on FGR. Sample size requirements were met in just 23% of studies, revealing systematic under-powering. Adherence to TRIPOD+AI guidelines varied, with consistent shortcomings in addressing model fairness, heterogeneity, and calibration, raising concerns about the generalizability of the findings.

Conclusions: The review highlights the need for standardized outcome definitions and improved methodological rigor in ML studies on FGR/SGA. Addressing these issues is essential for enhancing predictive models' reliability, generalizability, and clinical applicability in prenatal and postnatal care.

References:
1. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. doi:10.1136/bmj-2023-078378
2. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. BMJ. 2020;368:m441. doi:10.1136/bmj.m441.

P16. Correcting for differential underdiagnosis across protected attributes in clinical prediction models using cancer stage information and causal inference

Jose Benitez-Aurioles[1], Matthew Sperrin[1]
[1]*Division of Informatics, Imaging & Data Sciences, University Of Manchester*

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

Objective: We propose a novel method to correct for differential underdiagnosis in cancer prediction models.

Methods: In epidemiology, underdiagnosis in cancer is often indirectly measured through diagnostic delay, as some underserved groups are sicker at the time of diagnosis. If there are quantitative markers of disease progression, these could be used in order to understand which groups are diagnosed later, and correct for this. We show that this is possible, in the specific case of cancer stage, assuming that all people with late-stage cancer have the same probability of being diagnosed. We take a causal longitudinal approach (Figure 1), defining our estimand as the counterfactual patient-level risk of being diagnosed in a world in which diagnosis is not affected by patient characteristics.

Results: We provide theoretical proofs of the identifiability of these counterfactual predictions, and use a simulation to evaluate the method and benchmark it against alternative approaches.

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

P17. Missing confounding information in counterfactual prediction models: an example of model-based clinical evaluation in comparison of radiotherapy techniques in cancer

Jungyeon Choi[1], Artuur Leeuwenberg[1], Lotta Meijerink[1], Dr. Johannes Langendijk[2], Judith van Loon[3], Remi Nout[4], Johannes Reitsma[1], Karel Moons[1], Ewoud Schuit[1]
[1]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, [2]Department of Radiation Oncology, University Medical Center Groningen, [3]Department of Radiation Oncology (MAASTRO), [4]Department of Radiotherapy, Erasmus Medical Center Cancer Institute

Background: Counterfactual prediction enables comparisons of treatment effects between hypothetical alternative treatment options. In the Netherlands, patients with certain cancer types are selected for proton-based therapy over the predominant photon-based therapy based on predicted outcomes estimated from Normal Tissue Complication Probability (NTCP) models. With proton therapy being a relatively recent advancement, there is growing interest in estimating its causal benefit by evaluating what outcomes would have occurred had patients received photon therapy instead. Model-based clinical evaluation (MBCE) addresses this by contrasting the observed complications under proton therapy with counterfactual risks predicted by NTCP models for photon therapy.

Objectives: NTCP models, developed as prognostic tools, typically do not account for confounding factors in dose-outcome relationships, which are critical for causal effect estimation. This study investigates the impact of omitting confounding factors in NTCP models on the validity of MBCE estimates when comparing proton and photon therapy.

Methods: We simulated individuals eligible for photon therapy, varying the effect of a confounding factor on both radiation dose and complications after radiotherapy. Patient selection for proton therapy followed the procedure of National Indication Protocol for Proton therapy in the Netherlands. We estimated the average treatment effect (ATT) in the proton-treated group using both a minimal NTCP model (omitting the confounder) and a full NTCP model (including the confounder) and compared them with true treatment effects. In additional simulation settings, we modified patient selection by changing the predictor of the NTCP models.

Results: Both the minimal and the full model generally yielded unbiased ATT. While omitting the confounder reduced prediction accuracy of the minimal model, its ATT estimates remained unbiased. However, greater issues arose when variables associated with the patient selection process were excluded. Specifically, when predictors relevant to patient selection were omitted from the NTCP model used in MBCE, the resulting causal estimates were biased.

Conclusions: The simulations demonstrate that the counterfactual predictions from the NTCP models omitting confounding factors can still yield unbiased MBCE for treatment effect estimation between radiotherapy techniques. However, to enhance validity, NTCP models should incorporate variables that affects patient selection.

P18. A bivariate generalized linear mixed modeling approach to explore center effects in multicenter test accuracy studies

Jeremie Cohen[1], Costance Dubois[1], Patrick Bossuyt[2]
[1]Centre for Research in Epidemiology and Statistics (Inserm UMR1153), [2]Department of Epidemiology and Data Science, Amsterdam University Medical Centres, University of Amsterdam

Sensitivity and specificity are common outcomes in test accuracy studies. In multicenter test accuracy studies, these accuracy outcomes can vary across centers, leading to potential center effects that, if unaccounted for, may result in invalid accuracy estimates. Despite the importance of addressing these variations, there is currently no clear guidance on the appropriate statistical methods to account for center effects in test accuracy studies. In this work, we discuss the application of a bivariate generalized linear mixed modeling (GLMM) approach, which jointly models sensitivities and specificities in multicenter test accuracy studies using the exact binomial distribution. We illustrate this method through two case studies, demonstrating how it can be employed to explore center effects and address comparative accuracy questions by incorporating covariates into the model.

P19. Individual participant data meta-analysis to examine non-linear treatment-covariate interactions at multiple time-points for a continuous outcome

Miriam Hattle[1,2], Joie Ensor[1,2], Katie Scandrett[1,2], Marienke van Middelkoop[3], Danielle A. van der Windt[2,4], Melanie A. Holden[4], Richard D. Riley[1,2]

[1]National Institute for Health and Care Research (NIHR) Birmingham Biomedical Research Centre, [2]Institute of Applied Health Research, College of Medicine and Health, University of Birmingham, [3]Department of General Practice, Erasmus MC Medical University Center, [4]School of Medicine, Keele University

Background: Individual participant data (IPD) meta-analysis projects obtain, harmonise, and synthesise original data from multiple studies. Many IPD meta-analyses of randomised trials are initiated to identify treatment effect modifiers at the individual level, thus requiring statistical modelling of interactions between treatment effect and participant-level covariates. Using a two-stage approach, the interaction is estimated in each trial separately and combined in a meta-analysis. In practice, two complications often arise with continuous outcomes: examining non-linear relationships for continuous covariates and dealing with multiple time-points.

Objectives: We propose a two-stage multivariate IPD meta-analysis approach that summarises non-linear treatment-covariate interaction functions at multiple time-points for continuous outcomes. This allows for participants or trials with missing time-points to still be included in the analysis.

Methods: A set-up phase is required to identify a small set of time-points; relevant knot positions for a spline function, at identical locations in each trial; and a common reference group for each covariate. Crucially, the multivariate approach can include participants or trials with missing outcomes at some time-points. In the first stage, restricted cubic spline functions are fitted and their interaction with each discrete time-point is estimated in each trial separately. In the second stage, the parameter estimates defining these multiple interaction functions are jointly synthesised in a multivariate random-effects meta-analysis model accounting for within-trial and across-trial correlation. These meta-analysis estimates define the summary non-linear interactions at each time-point, which can be displayed graphically alongside confidence intervals.

Results: In this presentation, the proposed method will be described and demonstrated in detail. Additionally, the approach will be illustrated using an IPD meta-analysis examining effect modifiers for exercise interventions in osteoarthritis.

Conclusions: Modelling continuous covariates as non-linear allows for covariates to be modelled properly and avoid arbitrary categorisation into two or more groups. The illustrated example in the presentation will show evidence of non-linear relationships and small gains in precision by analysing all time-points jointly. Most trials provided all time-points, yet differences still arose. In situations with more missing time-points across studies, the gain in information will be more pronounced.

P20. Quantifying the versatility of routinely measured prognostic factors

Hamish Innes[1], Philip Johnson[2]
[1]Glasgow Caledonian University, [2]University of Liverpool

Background: Age is a versatile prognostic factor (PF) because it is able to predict diverse health outcomes. In this study, we sought to quantify this versatility and that of other commonly measured PFs.

Methods: This study was undertaken using the UK Biobank (UKB) cohort. Twenty continuous PFs, commonly used in routine clinical practice, were selected. These PFs included a range of renal, liver function, metabolic, lipid, blood count and other types of tests. All PFs selected were continuous numeric variables. Participants were followed from their UKB enrolment date (time zero) until date of censoring or the outcome of interest. More than 800 adverse health outcomes were considered, each corresponding to a specific 3-digit ICD code (e.g. A00, A01, A02, etc through to N97, N98 and N99). Outcomes with fewer than 20 events were omitted. Cox regression was used to determine the association between each PF with time to hospital admission for each outcome. All PFs were Z-transformed to ensure comparable log hazard ratios. Statistical significance was defined as $p<0.05$ with Bonferonni correction. The number of statistically significant associations, direction of the association (positive vs negative) and the median absolute log hazard ratio (LHR) were determined for each PF. Data were visualised using Volcano and Manhattan plots.

Results: The analysis included up to 502,408 UKB participants, followed for a median 12.4 years. PFs with the greatest number of statistically significant associations were age (563/836; median LHR: 0.47); glycated haemoglobin (478/831; median LHR: 0.10); hand grip strength (417/836; median LHR: 0.26); and albumin (412/828; median LHR: 0.23). PFs with the lowest number of significant associations were platelet count (232/833; median LHR: 0.16); bilirubin (199/833; median LHR: 0.17) and total protein (121/828; median LHR: 0.09) (Table 1).

Conclusion: Our study explores the concept of PF versatility, which has not been widely described hitherto. We confirm that age is singular in terms of its versatility, but also highlight additional PFs that are highly versatile (e.g. glycated haemoglobin, urea, handgrip strength and albumin). Understanding this versatility may inform PF selection strategies for prognostic clinical prediction models (CPMs) –e.g. inclusion by default of highly versatile PFs.

P21. Detecting violation of the conditional independence assumption using residual correlations in latent class models for diagnostic test evaluation: A simulation study

Yasin Okkaoglu[1], Nicky J Welton[1], Hayley E Jones[1]
[1]*Population Health Sciences, Bristol Medical School, University of Bristol*

Background: Latent class models have been adopted to estimate diagnostic test accuracy when there is no gold standard test available. To avoid biased estimates, it is crucial to capture dependencies between different tests within disease states. Residual correlation plots are often used to detect a lack of global fit of the model and to guide the selection of pairwise dependency terms for inclusion. We conducted an investigation to determine if residual correlation plots could effectively serve as a tool for detecting a lack of overall fit and for specifying the correct model across a broad range of scenarios.

Methods: We performed a simulation study to evaluate the performance of residual correlation plots in detecting deviations from the conditional independence assumption and identifying truly correlated test pairs. We generated 1000 binary test results for 4 diagnostic tests from a conditional dependence model where the first two tests are correlated within the diseased group. We examined the percentage of the time where (i) residual correlation plots identified an overall lack of fit and (ii) identified the correct correlated pair across 504 scenarios (3 sample sizes, 2 prevalences, 2 covariances, 42 sensitivity-specificity combinations).

Results: Failure to account for the conditional dependence between test 1 and test 2 led to over-estimation of the sensitivity of test 1 and test 2 (median: 0.09, 2.5th – 97.5th percentiles: 0.00 – 0.40), and under-estimation of prevalence (median: -0.05, 2.5th – 97.5th percentiles: -0.24 – 0.03). The residual correlation plots performed well in detecting a lack of global overall fit when sample size, prevalence and the extent of conditional dependence (covariance) were high. However, these plots only highlighted the correct pair (test 1 and test 2) as being the source of lack of fit 12.1% of the time, while suggesting a lack of pairwise fit for test 3 and test 4 64.9% of the time.

Conclusion: Relying exclusively on residual correlation plots for the model specification is not advisable. Instead, leveraging existing prior knowledge regarding the dependency structure among diagnostic tests or employing a comparative model selection strategy to choose among multiple models may be necessary.

P22. Comparison of methods to handle missing values in the index test in a diagnostic accuracy study – two simulation studies

Katharina Stahlmann[1], Dennis Juljugin[1], Bastiaan Kellerhuis[2], Johannes B. Reitsma[2], Nandini Dendukuri[3], Antonia Zapf[1]
[1]Institute of Medical Biometry and Epidemiology, University Medical Center Hamburg-Eppendorf, [2]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, [3]Department of Medicine, McGill University

Background: Most diagnostic accuracy studies apply a complete case analysis (CCA) or single imputation to address missing values in the index test, which may lead to biased results.

Objective: Two simulation studies were conducted to compare the performance of different methods in estimating the AUC of a continuous index test (study 1) and the sensitivity and specificity of a binary index test (study 2) given missing values in the index test.

Methods: We simulated data for a reference standard, index test, and three covariates using different sample sizes, prevalences of the target condition, correlations between index test and covariates (only study 1), and true AUC/sensitivity and specificity. Subsequently, missing values were induced for the index test, varying proportions of missing values and missingness mechanisms. Seven and six methods (study 1: multiple imputation (MI), empirical likelihood, and inverse probability weighting approaches; study 2: single imputation, MI, product multinomial framework-based methods) were compared to CCA regarding their performance to estimate the true AUC or sensitivity and specificity, respectively.

Results: Regarding the AUC under missing completely at random (MCAR) and many missing values, CCA gives good results with respect to bias for a small sample size and all methods perform well for a high sample size. If missing values are missing at random (MAR), all methods are severely biased if the sample size and prevalence are small. An augmented inverse probability weighting and standard MI methods perform well with higher prevalence and sample size, respectively.
Regarding sensitivity and specificity, MI outperforms all other methods under MCAR and MAR. Under missing not at random (MNAR), most methods in both studies are biased, with some improvements by higher correlation.

Conclusions: Most methods perform well given few missing values. For calculating the AUC under many missing values and MCAR or MAR, we recommend CCA given a small sample size and MI or the augmented inverse probability approach given a large sample size. For calculating sensitivity and specificity, MI should be used under MCAR and MAR. Regardless of outcome parameter, sensitivity analyses are recommended, especially if MNAR is likely.

Funding: Deutsche Forschungsgemeinschaft (DFG) grant number ZA 687/6-1

P23. Evaluating Treatment Benefit Predictors using Observational Data: Contending with Identification and Confounding Bias

Yuan Xia[1], Mohsen Sadatsafavi[1], Paul Gustafson[1]
[1]*University of British Columbia*

A treatment benefit predictor (TBP) maps patient characteristics into an estimate of the treatment benefit for that patient, which can support optimizing treatment decisions. However, evaluating the predictive performance of a TBP is challenging, as it often must be conducted in a sample where treatment assignment is not random. We show conceptually how to approach validating a pre-specified TBP using observational data from the target population, in the context of a binary treatment decision at a single time point. We exemplify with a particular measure of discrimination (the concentration of benefit index) and a particular measure of calibration (the moderate calibration curve). The population-level definitions of these metrics involve the latent (counterfactual) treatment benefit variable, but we show identification by re-expressing the respective estimands in terms of the distribution of observable data only. We also show that in the absence of full confounding control, bias propagates in a more complex manner than when targeting more commonly encountered estimands (such as the average treatment effect, or the average treatment effect amongst the treated). Our findings reveal the patterns of biases are often unpredictable and underscore the necessity of accounting for confounding factors when evaluating TBPs.

P24. Shapley Additive Explanations (SHAP) in healthcare research and predictive analytics: a scoping review and guide for tabular data applications.

Alex Carriero[1], Anne de Hond[1], Karel GM Moons[1], Maarten van Smeden[1]
[1]*University Medical Center Utrecht*

Background: The Shapley Additive Explanations (SHAP) framework is now ubiquitous in healthcare research and predictive analytics. SHAP is used to help circumvent the black-box nature of machine learning models. It is a post-hoc, model agnostic, explainable AI method, meaning that it can be used to generate model explanations for any pre-trained machine learning model, making it an attractive choice for researchers.

Methods: This article undertakes a detailed scoping review of 100 publications from 2023-2024 that reported using SHAP in their clinical research. We focused on publications that used (supervised) machine learning methods to model patient-related health outcomes using tabular data with clinical features. We provide an overview of i) how the SHAP explanations were calculated, reported and visualized ii) what SHAP explanations were used for and iii) highlight the prevalence of potentially misleading interpretations.

Results: We found that incomplete reporting with respect to SHAP methods was the norm. To aid in improved implementation and reporting of SHAP methods, this article offers a gentle introduction to the SHAP framework and provides concrete recommendations for reporting and interpretation.

Conclusion: The SHAP framework can be a valuable asset in healthcare research and predictive analytics, yet, without proper context to accompany the model explanations (complete reporting) their value can be obscured by ambiguity. We provide reporting recommendations for healthcare research involving SHAP and highlight how over-interpretation of explanations can be misleading.

P25. Validation of the summarization quality of GPT-generated discharge letters for routine hospital care

Anne De Hond[1], Laura Veerhoek[1], Ruben Peters[1], Tuur Leeuwenberg[1], Maarten van Smeden[1], Ilse Kant[1]
[1]*Umc Utrecht*

BACKGROUND: Large Language Models like GPT hold immense potential for application in the healthcare domain. To realize this potential, validation is a crucial first step to assess the quality of Large Language Models in healthcare.

OBJECTIVE: This study performed a thorough validation procedure to assess the quality of GPT-generated discharge letters based on Dutch patient records.

METHODS: Two senior medical students annotated the number of omissions, trivial facts and hallucinations/additions in 44 physician- and 44 GPT-generated discharge letters. Physicians evaluated the usability of the physician- and GPT-generated letters on a 5-point Likert scale and ranked both letters on their coherence, relevance, and overall quality.

RESULTS: Physician letters most often contained additions (53% of all minor and severe annotations), whereas the GPT letters most often contained omissions (55% of all minor and severe annotations). There were on average 1.7 omissions, 0 trivial facts, and 1.9 additions per physician letter. For GPT letters there were on average 3.6 omissions, 2.1 trivial facts, and 0.9 hallucinations. Physicians gave comparable ratings for the usefulness of GPT-generated and physician written discharge letters (Figure 1). They slightly preferred physician letters for their relevance and GPT letters for their structure. There was no clear overall preference for physician- or GPT-generated letters.

CONCLUSIONS: GPT-generated discharge letters are a promising tool for reducing administrative burden for routine clinical care. Our results underscore the importance of validation, as the presence of omissions, trivial facts, and hallucinations may impact downstream care processes. More research is needed on the clinical utility of GPT-generated discharge letters and the risks associated with hallucinations and other errors when applying GPT-generated discharge letters in clinical practice.

P26. Visualizing SHAP Values Associated with Hallucinations in Language Models

Ruurd Kuiper[1], Alex Carriero[1], Maarten van Smeden[1]
[1]*Julius Center, UMC Utrecht*

Background: Large language models (LLMs) have been successfully applied to many use cases, including text summarization. However, these models are prone to hallucinations—outputs that include incorrect or unsupported information. Especially in cases where accuracy is vital, such as healthcare, it is important to gain insight into when and why these hallucinations happen.

Objectives: This study aims to identify hallucinations in LLM-generated summaries and compute and visualize SHapley Additive exPlanations (SHAP) values associated with them. We hypothesize that tokens associated with hallucinations may exhibit distinct SHAP-value patterns, which might enable us to predict hallucinations in summaries where accuracy is not yet determined.

Methods: We used 100 samples from the XSum dataset [1], summarized using the Llama 3.2-1B model [2]. SHAP-values were calculated for each summary to measure the contribution of individual input tokens to the probabilities of each output token. Summaries were annotated for hallucinations using GPT-4 [3], which identified whether hallucinations were present, what they were, and which input tokens were most associated with them. A visualization tool was developed to show the following information for each generated summary:

1. The original article, generated summary, reference summary, and GPT-4-generated evaluation for each dataset sample.
2. An interactive display of SHAP-values associated with each token in the generated summary. Tokens associated with hallucinations are highlighted for easy reference.
3. A graph showing the magnitude of the SHAP-values as connections between tokens in the input text and the generated summary in one overview.

Results: The tool is accessible online at https://shap-sum.streamlit.app/. Users can interact with the data to visualize input-output relationships, identify tokens associated with hallucinations, and assess how input tokens influence these outputs.

Conclusions: This tool provides an intuitive and interactive visualization of SHAP-values associated with hallucinations in LLM-generated summaries. We aim to develop a predictive model for hallucination associated tokens based on the SHAP-values interactions found through this study.

References:
1. Narayan, S. et al. (2018). Don't Give Me the Details, Just the Summary! EMNLP 2018, 1797–1807.
2. Dubey, A., et al. (2024). The Llama 3 Herd of Models. arXiv:2407.21783.
3. Achiam, J., et al. (2023). GPT-4 Technical Report. arXiv:2303.08774.

P27. Monitoring AI-assisted screening performance on patient care pathway in a multi-centre, prospective, clinical trial: Statistical control charts and data requirements.

Jude Holmes[1], Sue Mallett[1]
[1]*UCL*

Background: Standard deviation (sd) of data is used to inform control chart rules in post-market surveillance (PMS), and CUSUM charts often use 3sd as an alarm. There is little guidance on study designs, sample size and alert rules for control chart and CUSUM monitoring, although within trial and PMS of tests is an important component of new diagnostic evaluation requirements from Regulators.

Objectives: To use statistical models to simulate monitoring the performance of AI during an RCT of breast mammography screening. We aim to evaluate alternative statistical strategies across a range of real-world inspired scenarios.

Methods: We will develop a model to simulate monitoring performance measured by arbitration rate, recall rate and cancer detection rates during an RCT of breast mammography screening. Our models will be informed by real world data on observed variations in recall and detection rates between centres and across calendar years.

Results: This project is part of an PhD which started in October 2024. We plan to report progress and results at the conference.

Conclusions: Clinically safe AI pathways are needed to bridge the gap from research to clinical use and must be guided by regulatory checks and balances. Using the described simulation to determine guidance on threshold recalibration is one step on this pathway.

P28. Evidence supporting angular velocity of the ankle during the gait cycle as a digital clinical outcome assessment (dCOA)

Helen Dawes[1], Eren Timurtas[2], Natasha Hassija[2], Sarah Donkers[3], Nancy Mayo[2]
[1]University Of Exeter, [2]McGill University, [3]University of Saskatchewan

Wearable technologies have revolutionized gait assessment moving this out of the laboratory and into the real world allowing for data acquisition on a quasi-continuous basis. Wearables are poised to augment data from standard clinical outcome assessments (COA) which can only be done periodically and in clinical settings yielding data that does not necessarily reflect the person's capacity for everyday activities nor indicate the treatable root causes of incapacity.

Researchers from McGill University, Canada, Exeter University UK, and PhysioBiometrics Inc. have developed a novel, smart, ankle-worn, wearable that provides kinematic data for every step taken during a walking session. The Heel2Toe™ wearable has been used in numerous clinical studies yielding an extensive data bank of kinematic and temporal spatial gait data.

Objective: The purpose of this study is to contribute evidence that signals extracted from angular velocities (AV) of the ankle during the different phases of the gait cycle relate to a standard COA, the Six Minute Walk Test (6MWT), supporting these metrics as dCOAs.

Methods: Data from Heel2Toe™ and standard COAs were available from people with Parkinson Disease (PWPD; n=23), and people with Multiple Sclerosis (PWMS; n=74) cross-sectionally and over time. Correlations were estimated between metrics from Heel2Toe™ and the 6MWT recorded simultaneously. Path analysis was used to link in the effect of non-motor symptoms.

Results: For PWPD and PWMS the strongest correlation was between average AV at push-off and the with absolute values ranging from 0.69 to 0.79. Path analysis showed that low mood affected both gait and function.

Discussion: AV generated during push-off was highly, even though not perfectly, correlated with 6MWT, a measure of functional walking capacity. However, AV metrics yield estimates of walking capacity more representative of the person's true capacity as well as identifying gait metrics to be targeted by therapy.

P29. Impact of and Reduction of Ancillary Features on CT and MRI LI-RADS Version 2018: An Individual Participant Data Meta-Analysis

Eric Lam[1], Danyaal H. Ansari[2], Kiret Dhindsa[3], Rebecca Thornhill[1], Christopher Sun[4], Haben Dawit[5], Christian B. van der Pol[6], Jean-Paul Salameh[7], Brooke Levis[8], Haresh Naringrekar[9], Hoda Osman[2], Mohammed Kashif Al-Ghita[2], Mostafa Alabousi[6], Mustafa R. Bashir[10], Andreu F. Costa[11], Matthew DF. McInnes[1,2,7]

[1]Ottawa Hospital Research Institute, [2]University of Ottawa, Faculty of Medicine , [3]Berlin Institute of Health at Charité, Universitätsmedizin Berlin, Berlin, Germany, [4]University of Ottawa Heart Institute, [5]Temerty Faculty of Medicine, University of Toronto, [6]Juravinski Hospital, and Cancer Centre, Hamilton Health Sciences, McMaster University, [7]Department of Radiology, University of Ottawa, [8]Centre for Clinical Epidemiology, Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Quebec, [9]Thomas Jefferson University Hospital , [10]Department of Radiology, Duke University Medical Center, Durham, NC, USA; Center for Advanced Magnetic Resonance Development, Duke University Medical Center, [11]Department of Diagnostic Radiology, Queen Elizabeth II Health Sciences Centre and Dalhousie University

Background: HCC is the leading cause of liver cancer and ranks as the third leading cause of cancer-related deaths globally. The Liver Imaging Reporting and Data System (LI-RADS) provides a standardized framework for diagnosing HCC using imaging features categorized on a 1–5 scale based on malignancy likelihood. While the diagnostic utility of LI-RADS major features is well-established, the contribution of ancillary features (AFs) remains unclear. Prior studies suggest that some AFs may have limited diagnostic utility, leading to potential inefficiencies and interobserver variability.

Objectives: This study aims to apply decision tree and machine learning methods to determine whether specific AFs are non-contributory to the diagnostic performance of the LI-RADS system. Specifically, the study seeks to identify AFs that may be non-contributory to diagnosis and validate prior findings from single-center studies within a multi-institutional database.

Methods: An IPD meta-analysis of 53 eligible studies, included in the LI-RADS Living Systematic Review (https://osf.io/tdv7j/) database, will be conducted. Eligible studies will evaluate LI-RADS major and ancillary features in patients at high risk for HCC using CT or MRI imaging. Diagnostic accuracy metrics, including area under the curve (AUC), sensitivity, specificity, and positive predictive value (PPV), will be calculated with and without the application of AFs. Decision tree analysis and machine learning methods, using logistic regression with L1 regularization, will be used to identify and validate non-contributory AFs. Model training and validation will be performed using a Stratified K-Fold iterator variant with non-overlapping groups cross-validation framework, where each study will appear exactly once in the test set across all folds.

Results: Preliminary results from prior single-center studies suggest that 13 of 21 AFs can be removed without affecting diagnostic performance. This study will validate whether these AFs can be similarly excluded in a multi-institutional context.

Conclusions: This study will inform the potential refinement of the LI-RADS guidelines by evaluating the diagnostic utility of ancillary features within the LI-RADS. Streamlining AFs may improve system efficiency and reduce variability without impairing diagnostic accuracy. These findings aim to optimize the diagnostic performance of LI-RADS, contributing to improved patient outcomes and more accurate clinical decision-making.

P30. Using IMU data from a wearable ankle mounted sensor to electronically measure a standard clinical outcome assessment – the 6 Minute Walk Test: from COA to eCOA using the Heel2Toe™ wearable

Nancy Mayo[1], Sarah Donkers[2], Helen Dawes[3]
[1]McGill University, [2]University of Saskatchewan, [3]University Of Exeter

Background: Electronically captured clinical outcome assessments (eCOA) can improve the efficiency of clinical trials and epidemiological studies and expand the type of outcomes assessed. Walking-related outcomes, such as distance walked in six minutes (6MWT), are highly relevant. PhysioBiometrics Inc. has developed a smart, shoe-mounted, training device, Heel2Toe™, that also assesses temporo-spatial and kinematic gait parameters from angular velocity (AV) of the ankle during walking. 6MWTs obtained outside the clinic and more frequently would better reflect the person's usual walking capacity.

Objective: To demonstrate a formula for estimating the 6MWT from data collected using the Heel2Toe™ sensor and illustrate the value of this e-COA using 12 case studies from the intervention arm of a trial of the therapeutic benefit of in-home gait training using the Heel2Toe™.

Methods: Over a 3-month period, 1089 walking sessions were recorded, 969 of 2 minutes or more. Three in-clinic assessments of 6MWT per person were planned and 32 of these 36 were carried out. An e-6MWT was estimated from clinic measured stride length (in meters) and time and from session-specific stride time, number of steps, and time spent walking.

Results: Clinic-6MWTs were higher than e-6MWTs for 17 of 32 available assessments. Six of 12 people (50%) showed improvement based on the clinic-6MWT and 4 of 12 (33.3%) using the e-6MWT. The numerous data points from the e-6MWT (6 to 143 per person) yielded 5 different shape patterns (step-improve, step-decline, U, J, flat) and the 3 data points for participants clinic-6MWTs yielded 6 different shape patterns (linear-improve, step-improve, step-decline, V, inverted-V, flat) but there was shape-agreement for only 3 of 12 participants: (one step-improve, one step-decline, one flat). Correlations between values of the e-6MWT and AV metrics varied across participants. e-6MWT correlated most often with heel strike AV and foot swing AV, 6/12 participants showing correlations between 0.30 and 0.74.

Discussion: Remote estimation of 6MWTs is feasible and provides data that differs from that of periodically measured 6MWTs. e-6MWTs were not related to gait metrics in some participants suggesting that changes for reasons for low walking capacity unrelated to kinematics need to be sought.

P31. IMPACT Framework: Bridging Gaps in Innovation and Advancing Health Technology

Katerina-Vanessa Savva[1], Rosario Luxardo[1], Melody Ni Zhifang[1], George B. Hanna[1], Christopher J. Peters[1]
[1]*Imperial College London*

Background: Health technologies have revolutionised healthcare by enhancing diagnostics, treatments, and preventive strategies. However, existing frameworks for evaluating their real-world impact often lack adaptability, comprehensiveness, and practical application, limiting their utility in guiding impactful research. There is a critical need for a robust, standardised tool to assess the multifaceted impacts of health technologies and inform research trajectories toward clinical utility.

Methods: We aim to develop IMPACT (Innovative Metrics for Progress and Clinical Translation) Framework, a comprehensive Impact Assessment Tool for Health Technologies through a multi-phase approach. First, a systematic literature review was conducted using Medline and Embase to identify existing tools and methodologies for evaluating health technology impact. Second, semi-structured interviews with key stakeholders (clinicians, policymakers, industry, patient representatives) were undertaken to capture diverse perspectives. Finally, a Delphi survey will be employed to achieve consensus on key dimensions and criteria for impact assessment. Real-life case studies will validate the tool, applying it to ongoing health technology projects to assess its reliability and practical utility. This case study will focus on a digital cancer surgery prehabilitation application called Onko.

Results: The development of this tool through a systematic mixed-methods approach, integrating insights from literature, expert interviews, and consensus-building via the Delphi survey, ensures a comprehensive and user-centred framework. By incorporating human perspectives, the tool addresses the nuanced realities of health technology adoption, including clinical, economic, and societal factors. This inclusive methodology enhances its relevance and adaptability across diverse healthcare settings. The tool is anticipated to identify critical research gaps, inform study designs, and direct research trajectories toward innovations with higher clinical impact and usability. Ultimately, it will empower stakeholders to align health technology development with real-world needs, fostering patient-centered and system-relevant solutions.

Conclusion: The IMPACT Framework will address limitations of existing frameworks by integrating diverse impact dimensions and enabling researchers to design studies with a clearer focus on clinical relevance. By identifying barriers and research gaps, the tool will promote the development of innovations with greater real-world applicability, optimising resource allocation and enhancing patient outcomes, from an early stage. This effort supports evidence-based decision-making and fosters the translation of health technologies from research to practice.

P32. Observational Cohort Study Describing the Distribution of Ferritin Levels and the Prevalence of Iron Deficiency in Patients with Chronic Conditions: A Retrospective Analysis Using CPRD Aurum

Alvin Katumba[1], **Cynthia Wright Drakesmith**[1], Suzanne Maynard[1], Sarah Haynes[1], Vijay Maharajan[1], Innocent Erone[1], Margaret Smith[1], Akshay Shah[1], Noemi Roy[1], Joseph Lee[1], Katja Maurer[1], Simon Stanworth[1], Clare Bankhead[1]
[1]Nuffield Department of Primary Care Health Sciences, University Of Oxford

Background: Iron deficiency is a common and clinically significant condition, particularly in patients with long-term medical conditions (LTMCs). Ferritin, a key biomarker of iron status, is crucial for diagnosing iron deficiency anaemia (IDA). However, its interpretation is often confounded by inflammation; which frequently observed in patients with LTMCs. To address this, alternative reference ranges for "normal" ferritin levels have been proposed for various chronic disorders, however further research is needed to identify the best approach for diagnosing and managing iron deficiency in these patients.

Methods: This retrospective observational study utilized anonymized patient data from the Clinical Practice Research Datalink (CPRD) Aurum database (February 2022 release), containing records from 41 million UK patients. The study described ferritin testing patterns, variability in ferritin levels, and the prevalence of iron deficiency in patients with chronic disorders based on WHO and NICE criteria.

Results: The cohort included 3,043,110 adults with at least one LTMC, of whom 88% (n = 2,672,872) had at least one recorded ferritin test between January 2015 and December 2021. Among these, 65% (n = 1,740,782) were female. A total of 8,774,231 ferritin tests were performed, with 68% conducted on female patients (n = 5,958,941).

Preliminary findings revealed significant variability in ferritin levels across LTMCs. Patients with coeliac disease and Crohn's disease demonstrated lower ferritin levels and a higher prevalence of iron deficiency compared to other LTMC cohorts, irrespective of cut-off criteria (WHO or NICE). These findings underscore the strong association between these inflammatory conditions and iron deficiency, emphasizing the need for targeted diagnostic and therapeutic strategies for high-risk groups.

Conclusions: This study highlights the feasibility of large-scale analyses of ferritin levels and iron deficiency using routine data. It underscores the variability of ferritin levels across LTMCs and the limitations of current thresholds. Future research will aim to establish tailored reference ranges and diagnostic approaches to optimise iron deficiency management in patients with chronic conditions.

P33. Optimising Ferritin as a Biomarker for Iron Deficiency in Cirrhosis: Insights from a Large-Scale Primary Care Dataset

Suzanne Maynard[1,2], Cynthia Wright Drakesmith[3], Innocent Erone[3], Sarah Haynes[4], Joseph Lee[3], Vijay Maharajan[3], Katja Maurer[3], Alvin Katumba[3], Noemi BA Roy[5], Akshay Shah[2,6], Margaret Smith[3], Simon Stanworth[2,7], Clare Bankhead[3]

[1]Radcliffe Department of Medicine, [2]NIHR Blood and Transplant Research Unit in Data Driven Transfusion Practice, [3]Nuffield Department of Primary Care Health Sciences, [4]Nuffield Department of Women's and Reproductive Health, [5]Department of Haematology, [6]Nuffield Department of Clinical Neurosciences, [7]NHS Blood and Transplant

Anaemia affects up to 80% of patients with cirrhosis and is associated with poor clinical outcomes. Iron deficiency, a common and modifiable cause of anaemia, may be underdiagnosed in this population. While ferritin <15 µg/L indicates absent iron stores, elevated levels due to chronic inflammation in cirrhosis complicate interpretation. This study evaluates ferritin as a biomarker for iron deficiency in cirrhosis to refine its clinical utility.

Using data from Clinical Practice Research Datalink (CPRD) Aurum, which includes anonymised records from 41 million UK patients, we describe testing patterns and diagnoses of iron deficiency anaemia (IDA), and initiation of oral iron therapy. The primary aim is to assess ferritin testing in cirrhosis. Secondary objectives include evaluating Hb response (≥10 g/L increase within 90 days of treatment) and ferritin's predictive value for treatment outcomes.

The cohort included 44,075 adults with cirrhosis and at least one Hb result (Dec 2015–Jan 2022). Ferritin was tested in 24,126 (55%) patients; 99% of results had a corresponding Hb result within 90 days. Among these patients, 63% (n=14,993) suffered from anaemia during the study. Using ferritin thresholds <15, <45, and <70, iron deficiency was identified in 19% (n=2907), 47% (n=7097), and 56% (n=8470) of anaemic patients, respectively. Of all patients with anaemia and a ferritin test during the study, 53% (n=7922) received oral iron prescriptions.

Multivariable regression models will explore ferritin's predictive value for Hb response, adjusting for confounders such as age, gender, ethnicity, long-term conditions, and inflammation (C-reactive protein). Sensitivity analyses will exclude patients recently treated with oral iron or those undergoing liver transplantation.

Preliminary findings demonstrate the feasibility of large-scale IDA analyses using routine data and are expected to highlight limitations of current ferritin thresholds. Future research will focus on redefining ferritin cut-offs and leveraging additional biomarkers, such as individual mean corpuscular volume (MCV) trajectory, to improve IDA diagnosis. Machine learning approaches may further enhance predictive accuracy. This research aims to optimise iron deficiency management and improve outcomes for patients with cirrhosis.

Supported by NIHR Blood and Transplant Research Unit (NIHR203334).

P34. Methods for comparing one and two reader mammography screens as part of the breast cancer screening programme

Breanna Morrison[1], Alice Sitch[1], Karoline Freeman[2], Rosalind Given-Wilson[3], Matthew Wallis[4], Louise Wilkinson[5], Sarah Pinder[6], Julia Brettschneider[7], Jackie Walton[8], Malcolm Price[1], Sian Taylor-Phillips[3]

[1]Test and Prediction Group - University Of Birmingham, [2]Warwick Screening - University of Warwick, [3]St Georges Healthcare NHS Foundation Trust, [4]Cambridge Breast Unit - Cambridge University Hospitals NHS Foundation Trust, [5]Oxford Breast Imaging Centre - Oxford University Hospitals NHS Foundation Trust, [6]Breast Pathology - Kings College London, [7]Department of Statistics - University of Warwick, [8]NHS England

Background: The breast screening programme in England currently requires two mammogram readers to assess each image. Prior research suggested screens read by two mammogram readers yielded both higher cancer detection and lower false positive rates than screens read by one reader. However, these studies used screens read by two readers and using the decisions made by a single reader within a pair, they assume those would be the decisions made if they were truly a single reader.

Purpose: This study used 1994-1999 breast cancer screening data which included 6 million mammograms from nearly 4 million women to look at screens that were truly read by only a single reader and compared them to screens where there were two or more mammogram readers. We additionally were able to analyse a small sample of mammogram readers who had experience reading as both a single reader and as part of a pair. Results of this analysis were also compared to the literature which used a different classification of one and two reader screens.

Results: For 1994-1999 data, both incident and prevalent screens, two reader screens were less likely to recall a woman for further tests (incident: aOR 0.860, prevalent: aOR 0.823), less likely to make a false positive recall to assessment (incident: aOR 0.859, prevalent: aOR 0.823), but also less likely to find cancer (incident: aOR 0.901, prevalent: aOR 0.802).

When looking at comparisons within readers (who read as both a single reader and as a pair), we found similar trends with mammogram readers recalling less, having less false positives, and finding less cancers while reading as part of a pair for incident screens though the difference in cancer detection was not significant.

Conclusion: While this study has some limitations, we were able to look at true single reader screens rather than making assumptions about mammogram readers decisions, and we were able to compare decisions from individuals when reading alone or in a pair. This study shows that single readers tended to recall more women and this may let to an increase in both false positive recalls but also cancers detected at screen.

P35. Target trial emulation in routinely collected primary care data to compare different monitoring strategies for people with long term conditions

Katie Charlwood[1], Martha Elwenspoek[1], Jessica Watson[1], Jonathan Sterne[1], Penny Whiting[1]
[1]*University of Bristol*

Background: The evidence-base for optimal monitoring strategies is weak. Current practice is largely based on expert opinion and local protocols vary, which has led to substantial variation in test use within the UK.

Objectives: To emulate an RCT in routine data including people with long-term conditions who are regularly monitored, comparing the impact of regular testing (intervention) versus no testing (control) with a certain test on patient outcomes.

Methods: We used data from Clinical Practice Research Datalink (CPRD) linked to Hospital Episode Statistics (HES). The methods were developed on a cohort of patients with newly diagnosed type 2 diabetes mellitus (T2DM) who received regular monitoring. The intervention consisted of approximately yearly liver function testing (LFT) compared to no regular LFT. Patient outcomes included hospitalisation and mortality. Patients who deviated from their assigned strategy were censored. We estimated the effect of LFT on each outcome using pooled logistic regression, weighted using stabilised inverse probability of censoring weights, controlling for baseline covariates.

Results: Of 221k patients with T2DM, 32k patients were eligible for trial 'recruitment'. At the start of the trial, 12,165 patients followed the control testing strategy (no LFT) and 19,951 patients the intervention strategy (regular LFT). Baseline characteristics, including age, sex, BMI, and indicators for comorbidity, healthcare usage, alcohol consumption, smoking, and deprivation, were well balanced. Maximum follow-up time was 23.3 months. Patients in the intervention arm were censored quicker than in the control arm. The main reason for censoring in both arms was switching testing strategy, e.g. people in the control arm started to receive LFT. We found an increased risk of hospitalisation in people who received regular LFT (OR 1.08, 95% CI 1.03 – 1.14), which may be due to unmeasured confounding, but no difference in all-cause mortality (1.27, 0.99 – 1.66).

Conclusions: These findings suggest that regular LFT does not impact all-cause mortality within 23 months after T2DM diagnosis, but may increase the likelihood of hospitalisation. However, due to significant limitations these results should not be used in isolation to change monitoring practices. Longer follow-up and better coding of tests would improve the utility of these methods.

P36. Prioritisation of Unmet Diagnostic Needs in Care Homes through an online, modified Delphi survey.

Tim Hicks[1], Sara Pretorius[1], Jana Suklan[1], Louise Jones[2], Dave Belshaw[3], Miles Witham[1]
[1]NIHR Healthtech Research Centre In Diagnostics And Technology Evaluation, [2]Northumbria Healthcare NHS Foundation Trust, [3]Health Innovation North East and North Cumbria

Background: The COVID-19 pandemic pushed community diagnostics to the forefront of healthcare, with lateral flow tests being conducted daily. One of the highest profile settings to use such a diagnostic was that of a Care Home, which often have some of the most vulnerable members of society. Following the pandemic, there is an increased desire to decentralise diagnostics, moving away from primary and secondary care, and promoting more community-based approaches. To do this efficiently, we need to understand where there are gaps (termed unmet needs), and where developers, commissioners, and healthcare staff should focus their efforts.

Objectives: This study aimed to understand the priority for new diagnostics in a care home setting from the perspectives of healthcare staff involved in a care home setting, and to generate consensus around the highest priority needs to enable rapid literature reviews for available technology.

Methods: Following an initial round of interviews with healthcare staff (n = 16) to elicit unmet diagnostic needs, we disseminated a two-round modified Delphi survey to participants (n = 44) who ranked the priority for new care home diagnostics across five themes; Acute Deterioration, Cardiovascular Health, Chronic Conditions, Infection, and Other Diagnostics. Participants included a mixture of Care Home Managers, Nurses, General Practitioners, and Geriatricians (see Figure 1). To generate consensus, participants were shown the group results of the first round to allow them to adjust their answers if they desired.

Results: The five areas were prioritised (see Figure 2) with strong consensus from the participants showing minimal change between rounds. The 40 Unmet diagnostic needs within these areas showed stability of consensus, with the highest priority unmet needs solidifying the consensus between rounds (see Figure 3 for top unmet needs in each area).

Conclusions: There are a substantial number of unmet diagnostic needs in care homes, with strong agreement between healthcare staff as to their priority. Following this work, two of the highest rated unmet diagnostic needs within the Infection theme (diagnostics for Urinary Tract Infections and Respiratory Tract Infections) were taken forward for a rapid review of upcoming diagnostics to try and address these gaps.

P37. PROVIDENT Imaging guidance (PROspectiVe Imaging DEsign and coNduct for Trials): Avoiding common pitfalls in the design and conduct of imaging trials

Katie Biscombe[1], **Sue Mallett**[2], Nuria Porta[1], Tom Nicols[3], Liz Hensor[4]
[1]The Institute of Cancer Research, [2]Centre for Medical Imaging, University College London, [3]Big Data Institute, University of Oxford, [4]Leeds Institute of Rheumatic and Musculoskeletal Medicine, University of Leeds

Background: Imaging can be used for diagnosis, guiding treatment, assessing treatment response and monitoring disease but adds complexity to the study design, conduct and analysis of clinical trials and studies, including collecting, using and storing imaging data.

Objectives: We aimed to bring together multidisciplinary perspectives to develop guidance and recommendations for addressing challenges in conducting prospective clinical research studies that include imaging. This guidance will be useful for funding applications, protocol, trial implementation, conduct and commercialisation and uptake of new imaging techniques.

Methods: The NIHR Statistics Imaging Studies Working Group statisticians involved in prospective imaging trials brought together 7 multidisciplinary focus group meetings including 27 participants to identify different domains and the specific methodological challenges that imaging studies present in these areas. Multidisciplinary groups included radiologists, radiographers, clinicians, patient advocates, trial managers, health economists, research nurses, methodologists, statisticians, and central lab and site perspectives.

Results: We mapped out guidance on 49 items grouped into 12 domains: ethics, participant information and consent, recruitment, imaging acquisition and processing, imaging scoring logistics, data flow and storage, QA/QC, site set-up, training, trial conduct, health economics, and commercialisation. For each domain we clarify the importance of items linked to examples from real-world prospective studies.

Conclusions: Our guidance spanning 12 domains of prospective imaging studies will enable trialists to avoid pitfalls by benefitting from multidisciplinary experience from prior studies. Future work will address challenges in statistical analysis of imaging studies. Many prospective imaging studies could be improved by the upfront awareness of potential challenges and understanding of real-world examples provided in our guidance.

P38. The linked evidence approach in the Institute for Quality and Efficiency in Health Care (IQWiG): an analysis of benefit assessments of diagnostic and screening tests

Ulrike Paschen, Sebastian Grümer, Markus von Pluto Prondzinski
[1]*Institute for Quality and Efficiency in Health Care (IQWiG)*

Background: The evaluation of screening and diagnostic tests requires assessment of the impact of the complete test-treatment-strategy. However, studies on test-treatment-strategies are often lacking (direct evidence). In order to address this shortcoming linking of the evidence of the test-treatment-strategy components – test accuracy and treatment effectiveness – may fill the gap (linked evidence approach [LEA]).

Objectives: Here, we analysed how often LEA was considered in benefit assessments of diagnostic and screening tests at the IQWiG and how often linkage of the LEA-components was successfully undertaken (i. e. LEA-components were available and actually linked). Additionally, we investigated the reasons why LEA-components could not be linked.

Methods: All benefit assessments of diagnostic and screening tests at the IQWiG with a final report prepared between 2004 and 2024 were examined. The following criteria were considered: the kind of test (diagnostic or screening), LEA considered (yes or no), the conclusion regarding test benefit (effective or not effective), the level of certainty of the conclusion (low, moderate or high), the frequency of completed linkage, and the reasons for why linkage was not successfully undertaken. Additionally, in reports with successful linkage, the study design of the therapy component was extracted (randomization yes or no).

Results: Out of a total of 51 eligible reports (24 on diagnostic tests and 27 on screening tests) almost half considered LEA (n = 25), comprising 7 projects on diagnostic tests and 18 projects on screening tests. Linkage of the LEA-components was successfully undertaken in 6 benefit assessments (24%; 2 diagnostic tests and 4 screening tests), all concluding that the test was beneficial. In these 6 reports with successful linkage the study design of the therapy component was mainly non-randomized. The level of certainty of the conclusions ranged from low to high. The reasons for why LEA-components could not be linked concerned various elements of the test-treatment-strategy, the most common reason (9 out of 19) being unclear benefit of (earlier) treatment.

Conclusion: LEA is a valuable approach for benefit assessments of diagnostic and screening tests.

P39. Blood tests in children with fatigue in Dutch primary care: a retrospective cohort study based on a routine primary care database (AHON)

Iris Baars[1], Guus (CGH) Blok[1], Michiel (MR) de Boer[1], Huibert (H) Burger[1], Manfred (MM) Schweiger[1], Otto (OR) Maarsingh[2,3], Tim (T) olde Hartman[4], Gea (GA) Holtman[1]
[1]Department of Primary and Long-Term Care, University of Groningen, University Medical Center Groningen, [2]Department of General Practice, Amsterdam University Medical Center, Location Vrije Universiteit Amsterdam, [3]Amsterdam Public Health Research Institute, [4]Department of Primary and Community care, Research Institute for Medical Innovation, Radboud University

Background: Dutch GPs encounter approximately 15 children with persistent fatigue annually. Somatic diagnoses in these cases are rare, and the lack of clear guidelines for blood testing could lead to over- and under-testing, as well as variability in testing among GPs. Insight into current management of blood testing in fatigued children in primary care could help identify opportunities for improvement.

Objectives: To describe the current management of fatigued children in primary care, focusing on the use of diagnostic blood tests.

Methods: We conducted a retrospective cohort study using the AHON primary care database. Children aged 4–18 years who consulted their GP for ICPC-A04 (weakness/tiredness general) between 2015 and 2022, without any prior consultations for fatigue in the preceding year, were included. Data on patient characteristics, somatic diagnoses, blood tests, prescriptions, and referrals at first consultation were collected, and follow-up GP consultations were tracked during a one-year follow-up.

Results: Among 6,859 fatigued children (median age 14 years; IQR 10-15), 2,506 (36.5%) were male. Somatic diagnoses (e.g. diabetes and hypothyroidism) were found in 1,805 (26.3%) patients, but only 8 cases were linked to the fatigue episode by the GP. Blood tests were conducted at first consultation in 4,566 (66.6%) children, involving 3,343 unique test combinations. Common tests included Hb and a multi-test panel (ALAT, ESR, GGT, glucose, Hb, potassium, creatinine, leucocytes, MCV, sodium and TSH), performed in 171 (2.5%) and 89 (1.3%) children, respectively. The Dutch Primary Care Guideline (LESA), recommending testing ALAT, CRP or ESR, glucose, Hb, creatinine, and TSH, were followed in 488 (7.1%) children, but always as part of a broader panel. Abnormal test results were observed in 3,231 children (47.1%). Within two weeks of the initial consultation, 1,514 (22.1%) children received medication, 191 (2.8%) were referred to secondary care and 5,009 (75,5%) had at least one A04-related follow-up consultation within a year.

Conclusion: Despite the low incidence of somatic diagnoses potentially explaining fatigue (<1%), GPs frequently use blood testing to manage fatigued children, with considerable variation in testing strategies. This highlights the need to further investigate the cost-effectiveness of different blood testing approaches for children with fatigue.

P40. Diagnostic Prediction Models For Spinal Fractures: A Systematic Review With Meta-Analysis Of The Canadian C-Spine Rule

Daniel Feller[1], Roel Wingbermühle[2], Edwin Oei[3], Bart Koes[1], Alessandro Chiarotto[1]
[1]Department of General Practice, Erasmus MC, University Medical Center, Rotterdam, The Netherlands, [2]Department of Physiotherapy and Rehabilitation Sciences, SOMT University of Physiotherapy, Amersfoort, the Netherlands, [3]Department of Radiology & Nuclear Medicine, Erasmus MC, University Medical Center, Rotterdam, The Netherlands

Background: Spinal disorders are the leading cause of disability worldwide. While most spinal conditions are non-specific, some involve serious pathologies requiring emergency care, with spinal fractures being the most prevalent.

Objectives: To evaluate the predictive performance of multivariable diagnostic models for identifying spinal fractures in patients with spinal pain and/or trauma.

Methods: We prospectively registered the protocol of this systematic review in PROSPERO. Observational studies developing and/or externally validating diagnostic models for spinal fractures were included. We searched MEDLINE, EMBASE, and Web of Science up to May 2024, with the addition of backward and forward citation tracking strategies. Two independent reviewers screened studies and extracted data. We extracted data on the models using the CHARMS checklist and assessed their risk of bias using the PROBAST tool. We used a bivariate random-effects meta-analysis to pool the sensitivity and specificity of studies externally validating the Canadian C-spine Rule (CCR) and a univariate random-effects meta-analysis to pool the discrimination of the same studies.

Results: We included 27 studies reporting on 34 diagnostic models. All models had an overall high risk of bias, with applicability concerns arising from the frequent use of spinal injuries as outcomes rather than spinal fractures. Meta-analyses of ten studies validating the CCR in adults with trauma showed excellent sensitivity (0.999; 95% CI 0.976–1) and a good AUC (0.85; 95% CI 0.72–0.97). However, specificity was low (0.188; 95% CI 0.063–0.443), with a pooled non-significant positive likelihood ratio of 1.23 (95% CI 0.978–1.548) and a negative likelihood ratio of 0.007 (95% CI 0.001–0.082). Other models for traumatic cervical fractures and osteoporotic fractures showed promise but lacked external validation or sufficient reporting on calibration, discrimination, and clinical utility. None of the thoracic and/or lumbar fracture models are ready for clinical use.

Conclusion: The CCR is an effective screening tool for traumatic cervical fractures in emergency settings. However, we did not identify any externally validated models suitable for clinical use regarding osteoporotic fractures, traumatic fractures of the thoracic and/or lumbar spine, and traumatic fractures of the cervical spine in non-emergency settings. Future research with rigorous methodological approaches should aim to fill these gaps.

P41. QUADAS-3: updated tool to evaluate risk of bias and applicability concerns in diagnostic test accuracy studies

Alison Suhsun Liu[1,2], Margi Shah[3], Louis Leslie[1], Juien Lo[4], Ben Harnke[5], Scott Hauswirth[6], Gianni Vergili[7,8], Tianjing Li[1,2]
[1]Department of Ophthalmology, University Of Colorado Denver Anschutz Medical Campus, [2]Department of Epidemiology, University Of Colorado Denver Anschutz Medical Campus, [3]Independent researcher, [4]Department of Internal Medicine, Metrohealth Medical Center, [5]Strauss Health Sciences Library, University of Colorado Anschutz Medical Campus, [6]Dompé farmaceutici, S.p.A., [7]IRCCS-Fondazione Bietti, [8]NEUROFARBA, University of Florence

BACKGROUND: Meibomian gland dysfunction (MGD) is a leading etiology of dry eye, affecting an estimated one-fifth of the US general population. Traditional diagnostic work-up includes clinicians' subjective evaluation of meibography. Artificial intelligence (AI)-based image grading shows promises in assisting with screening or staging.

OBJECTIVES: To synthesis the diagnostic performance of AI-based meibography grading as compared with human graders.

METHODS: We followed the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy and performed searches across four databases in June 2024. We included studies enrolling participants of any age, sex, or ancestry, whether symptomatic or not, if they evaluated machine learning or deep learning models for MGD. To assess potential risk of bias and applicability, we used a modified Quality Assessment of Diagnostic Accuracy Studies 2 checklist. We applied bivariate logistic models to estimate summary sensitivity and specificity.

RESULTS: We identified 8 eligible studies (12 evaluations) involving 3886 predominantly middle aged (average age: 27 to 53 years) participants who were primarily female (≥ 54.5%). Meibography images were obtained through noncontact infrared (75%) or in vivo confocal microscopy. Most algorithm (92%) employed deep learning models; only 17% were validated using cross validation techniques. All studies had a high risk of bias in at least one domain, with 87% raising high concern about applicability. Based on 9 single-algorithm models, the estimated summary sensitivity and specificity was 92.5% (95% confidence interval [CI] 85.9%-96.1%) and 92.9% (95% CI 81.2%-97.6%), respectively. AI algorithms trained to distinguish MGD from a mixture of normal Meibomian glands and mild MGD or other MGD type had an increased sensitivity (96.8%, 95% CI 96.0%-97.5%) but decreased specificity (89.2%, 95% CI 87.3%-90.9%). Potential significant sources of heterogeneity included image type, eyelids examined, validation technique, and study design.

CONCLUSIONS: The current evidence on AI-based grading of meibography images has low certainty due to uncertainty in accuracy metrics, high risk of bias, and concerns about clinical applicability. To strengthen the evidence and support real-world use, future studies should adopt rigorous designs, include more unique images, and develop externally validated models.

P42. Do the assays fit the use-case? A systematic review of exhaled breath-based biomarkers for tuberculosis detection.

Emily MacLean[1], Alvin Kuo Jing Teo[1], Mary Gaeddert[2], Tess Lai[3], Mikashmi Kohli[4], Prof Greg J Fox[1], Dr Morten Ruhwald[4], Prof Claudia M Denkinger[2]
[1]The University Of Sydney, [2]Heidelberg University, [3]Centenary Institute, [4]FIND

Background: Approximately three million cases of tuberculosis (TB) are missed each year, partially due to difficulties with existing assays, such as test placement and sample requirements. Therefore, developing simple, integrated assays that can be deployed in decentralised settings may improve testing access. Additionally, available TB tests require sputum samples, which are often difficult to produce for key populations, including children and people living with HIV (PLHIV). Investigating alternative samples, such as exhaled breath, for TB detection is therefore of high interest.

Objective: To synthesise evidence on exhaled breath-based biomarkers for the diagnosis of pulmonary TB.

Methods: We searched three databases from 2005 to 2023. Studies reporting the performance of any exhaled breath collection technique used in combination with any biomarker detection assay for TB diagnosis were included; biomarkers of host or pathogen origin were acceptable. To capture performance in all age groups, microbiological and clinical reference standards were included. Two independent reviewers screened all citations and conducted data extraction. No meta-analyses were performed due to high heterogeneity between studies and assays.

Results: From 6277 unique citations, 43 studies were included. Studies described assays that were either (i) portable integrated instruments that both captured and read samples (eNoses'), or (ii) combinations of stand-alone breath collection devices (e.g., commercially available disposable bags or facemasks) with separate biomarker readers (e.g., molecular tests, immunoassays). Eleven studies described accuracy of portable eNoses, five examined Xpert MTB/RIF, with the remaining utilised assays requiring advanced laboratory infrastructure, e.g., gas chromatography-based methods. Two exclusively paediatric studies recruited a combined 34 participants, while 19/43 studies included PLHIV. Assay development phases ranged from initial discovery to later-phase prospective validation studies, with most studies reporting early verification of biomarker signals, devices, or both. Accuracy of breath-based TB assays tended to decrease as development phase advanced.

Conclusions: Although an area of growing research interest, the applicability and accuracy of current breath-based TB assays is sub-optimal. Further assay refinement to simplify kits so they may be placed in peripheral settings would improve their utility. More research among key populations who may benefit most from these assays is required.

P43. Stronger penalties on treatment-covariate interactions improve treatment effect predictions and prevent potential treatment mistargeting

Chikomborero Mutepfa[1,2,3], Gurdeep Sagoo[1,3], Kile Green[1,2,3]
[1]NIHR HRC Diagnostic and Technology Evaluation, [2]The Newcastle upon Tyne Hospitals NHS Foundation Trust, [3]Newcastle University

Background: Non-tuberculous mycobacterial pulmonary disease (NTM-PD) is a rare infection of the lungs caused by non-tuberculous mycobacteria, most commonly the subspecies Mycobacterium abscessus (MAB) and Mycobacterium Avium Complex (MAC). Diagnosis is based on clinical history, computed tomography findings, and positive findings from acid-fast bacilli (AFB) cultures. We aimed to determine the ideal scenario for introducing a rapid molecular NTM assay that has near perfect analytical sensitivity and specificity in detecting MAC and MAB.

Methods: We conducted a cost analysis comparing the molecular assay as a replacement test to AFB cultures for patients suspected to have MAC-PD or people with Cystic Fibrosis suspected to have MAB-PD from the UK NHS perspective. The time horizon was 18 months. A decision tree Markov model was developed in TreeAge Pro Healthcare, v2021 R2.1. One-way sensitivity analyses were conducted to identify cost drivers that could be useful in guiding further test development. The NTM assay was modelled at no cost in the analysis to determine pricing and where the test could be used.
Results: In the analysis of patients with suspected MAC-PD, the novel assay exhibited cost savings, with average total costs of £194 per patient compared to £395/patient in the AFB culture pathway. Cost savings were also gained using the assay in the scenario analysis of patients suspected with MAB-PD, (£886/patient compared to £2,262/patient). The main cost drivers for the model were the prevalence and the specificity of the molecular assay and the AFB culture. Increasing the prevalence increased the total costs in both the AFB culture and molecular assay pathways, while an increase in specificity decreased the total costs. The model estimated a price for the NTM assay of up to £103 or £706 for patients with suspected MAC-PD or MAB-PD, respectively.

Conclusions: The replacement of AFB culture with the molecular assay in diagnosing patients with NTM-PD could improve the treatment and management of NTM due to the better identification of MAC or MAB infections and produce cost savings. Albeit, given the paucity of data used in modelling, results may not accurately reflect the use of resources and outcomes in the real world.

P44. Identifying examiners with similar diagnostic notion – Secondary analysis of a diagnostic accuracy study with 61 examiners aiming to identify exposed dentin in eroded teeth

Kirstin Vach[1], Carolina Ganss[2], Nadine Schlueter[1], **Werner Vach**[3,4]
[1]Hannover Medical School, Department of Conservative Dentistry, Periodontology and Preventive Dentistry, [2]Department of Operative Dentistry, Endodontics, and Paediatric Dentistry, Section Cariology of Ageing, Philipps-University Marburg, [3]Basel Academy for Quality and Research in Medicine, [4]Department of Sport Sciences and Biomechanics, University of Southern Denmark

Background: For many diagnostic questions, the task to come to a correct diagnosis is not trivial and may include subjective elements. Consequently, the decision-making process of multiple examiners can be expected to produce heterogeneous results. Understanding the sources of heterogeneity can assist in improving the decision-making process and increasing the diagnostic accuracy.

Objectives: Using the diagnosis of exposed dentin in eroded teeth as an example, we hypothesized that there are subgroups of potential examiners who share a diagnostic notion and may differ in the way how visual cues are taken into account. We aimed at identifying examiners with a similar diagnostic notion in order to identify and understand existing notions.

Material and methods: The work is a secondary analysis of study in which 61 examiners (23 dentists, 18 scientists, 20 dental students) visually assessed 49 tooth surfaces with varying degrees of erosion in terms of dentin exposure or non-exposure. The actual status was determined histologically. For each area the severity of erosion and the background color was recorded. Gender, age, and professional status, experience and specialty of the examiners were documented as person-specific characteristics. A specifically designed algorithm was used to search for clusters of examiners who frequently agreed in their decisions.

Results: The cariology specialists participating in the study did not agree in their decisions. Four clusters of examiners with similar decisions could be identified. Examiners in the cluster with the highest diagnostic accuracy showed heterogeneous person-specific characteristics, in particular they covered all levels of experience and specialty. Compared to the other clusters, they oriented themselves least on the severity of the erosion and only to a limited extent on the tooth background color. A few teeth were rated very differently across the clusters, and some of them were characterized by visual abnormalities.

Conclusions: The described methodology can be helpful to identify clusters of examiners with similar decision-making processes which can assist in better understanding of explicit and implicit mechanisms of diagnostic decisions.

P45. Targeted test evaluation: Five suggestions for refining the framework for designing diagnostic accuracy studies with clear study hypotheses

Werner Vach[1,2]
[1]Basel Academy for Quality and Research In Medicine, [2]University of Southern Denmark

Five years ago, Korevaar and colleagues proposed a framework for designing diagnostic accuracy studies, focusing on the definition of clear study hypotheses. This proposal filled a gap and was well received by the scientific community. In this talk, I will suggest five potential refinements. They aim at increasing the flexibility of the framework while pertaining or improving its logical consistency.

1) The relationship between minimal criteria and the choice of the null hypothesis region can be made more explicit by defining a region of clinical uselessness and a region of clinical usefulness.

2) A potential compensation between sensitivity and specificit can be allowed by adjusting the shape of the region of clinical uselessness.

3) If weighing between false postive and false negative decisons in based on other parameters than sensitivity and specificity, they can be directly used to define the regions.

4) The analytical strategy of the study can be phrased as an estimation problem without any change in sample size considerations.

5) Directly moving to a comparative accuracy study may facilitate the phrasing of study hypotheses and minimal criteria.

P46. Quantifying the Clinical Usefulness of Novel Biomarkers and Tests: Insights from Scenario Analyses

Frank Doornkamp[1], Jelle Goeman[1], Ewout Steyerberg[1,2]
[1]LUMC, Leiden University Medical Center, [2]UMCU, University Medical Center Utrecht

Background: New prognostic tests and biomarkers continue to emerge, often characterized by their sensitivity and specificity. It is unclear how prognostic testing translates into better health benefits in clinical contexts.

Objectives: We aim to evaluate the clinical usefulness of new biomarkers by quantifying how they may improve health outcomes through improved treatment allocation, compared to current standard care.

Methods: We developed a decision analytic model to illustrate the clinical usefulness of a new biomarker. Risk distributions were simulated based on current standard risk assessment alone or with the inclusion of the new test. Individual treatment benefit was estimated assuming a constant relative effect. Treatment was recommended if the absolute risk reduction exceeded a defined treatment threshold. Treatment decisions based on current standard risk assessment were compared to decisions made while including the new test. The clinical usefulness of a new test was quantified with Net Benefit: a weighted sum between the reduction in events and the number of treatments given, per 10,000 patients. As an illustrative example, we assessed the clinical usefulness of the genomic MammaPrint test (assumed sensitivity 62%, specificity 68%) in addition to the standard clinical risk assessment (PREDICT model, https://breast.predict.cam/tool) for early breast cancer patients. Systematic sensitivity analyses assessed the drivers of clinical usefulness.

Results: For early breast cancer, the MammaPrint had clinical usefulness when added to the standard clinical risk assessment, improving Net Benefit from 13 net distant metastases prevented to 15 per 10,000. Sensitivity analysis considered a range of treatment thresholds. The clinical usefulness of a new test was larger if the event rate was higher, treatment effect larger, its own quality better, or quality of the reference model lower. We present test and context characteristics in a ShinyApp to facilitate early assessments of the potential clinical usefulness of novel tests and biomarkers.

Conclusions: Decision analytic modeling provides insights into how the sensitivity and specificity of a new biomarker translate to clinical usefulness within its clinical context. We found that clinical usefulness depends not only on its prognostic strength but also on key contextual factors.

P47. Patterns of temporal trends in diagnostic accuracy estimates from meta-analyses of studies published in the Cochrane database of systematic reviews

Jacqueline Murphy[1], Thomas Fanshawe[1]
[1]*Nuffield Department Of Primary Care Health Sciences, University Of Oxford*

Background: Guidance on postmarket surveillance includes recommendations to monitor changes in the performance of diagnostic devices post-approval. This study builds on existing research into methods for evaluating time trends in diagnostic test accuracy (DTA) studies in the context of meta-analysis.

Objectives: The objectives of this study were to evaluate the prevalence and patterns of temporal trends in meta-analyses within DTA systematic reviews registered with the Cochrane Database of Systematic Reviews (CDSR), and to assess the suitability of existing methods for analysing such trends.

Methods: We analysed data from meta-analyses within DTA systematic reviews registered with the CDSR that were published between 20th August 2016, and 20th August 2023. For each review we conducted cumulative bivariate random effects meta-analysis of sensitivity and specificity estimates after ranking studies by publication date. We described temporal trends graphically with plots of summary estimates by study rank, and with receiver operating characteristic curve plots. We analysed linear trends using weighted linear regression with autocorrelated errors of summary estimates against study rank.

Results: The analysis included 46 reviews (92 meta-analyses of sensitivity and specificity). The total number of studies within all reviews was 1,486 with a median (IQR) 7,134 (2,782–16,406) participants per review. Temporal trends in at least one DTA measure were observed in 40 (87%) reviews, and statistically significant linear trends in 32 (70%) reviews. Non-linear trends were observed in 14 (30%) reviews. In 6 (13%) reviews there was no evidence for a temporal trend in either DTA measure.

Conclusions: This study adds to existing research by characterising various patterns of both linear and non-linear temporal trends in DTA meta-analyses. Graphical methods for illustrating trends are discussed. The study illustrates the importance of considering temporal trends when reporting results of diagnostic accuracy meta-analyses, and checking for the presence of non-linear trends before using existing trend analysis methods.

P48. Blinded Sample Size Re-estimation in Comparative Diagnostic Accuracy Studies Taking into Account Missing Data

Mahnaz Badpa[1], Katharina Stahlmann[1], Antonia Zapf[1]
[1]*University Medical Center Hamburg-Eppendorf*

Background: Diagnostic accuracy studies evaluate how well tests identify the presence or absence of a condition. They provide metrics such as sensitivity, specificity, and AUC to guide clinical decisions. Sufficient accuracy is essential for certification and a prerequisite for improving patient outcomes (1,2). Accurate sample size calculation ensures that diagnostic studies have sufficient power to draw meaningful conclusions. However, traditional methods often rely on fixed assumptions, which can lead to underpowered or inefficient studies (3,4). Adaptive designs address these limitations by allowing pre-planned modifications during the study based on interim data. These designs improve resource use and maintain statistical validity. Blinded sample size re-estimation adjusts sample sizes during the study and requires no adjustment of the type I error (5). This study evaluates blinded sample size re-estimation in a comparative diagnostic accuracy study and its efficiency with different methods for handling missing data.

Methodology: We conducted a simulation study to evaluate sample size re-estimation based on the prevalence in a paired framework comparing two tests. Initial sample sizes were calculated using predefined assumptions for disease prevalence and expected AUC values. Scenarios varied key parameters including disease prevalence, AUC, missing data proportion, and missingness mechanism. In total, 396 scenarios were created, each being simulated 1,000 times. Blinded sample size re-estimation was conducted at a pre-specified interim point when 50% of data was collected. Methods for handling missing data included complete case analysis (CCA), Empirical Likelihood-based Hot Deck (HDEL), and multiple imputation (MI).

Results: The adaptive design adjusted sample sizes to align with true prevalence, maintaining statistical power and efficiency across the majority of scenarios. When true prevalence exceeded assumptions, the adaptive design increased sample sizes to maintain power, and reduced them for overestimated prevalence to avoid over-recruitment. The adaptive design consistently produced less biased and more robust accuracy estimates compared to the fixed design. MI showed the lowest bias across all scenarios, whereas CCA exhibited significant bias, especially with higher proportions of missing data.

Conclusion: The application of adaptive sample size provides a flexible and efficient approach to managing uncertainties in diagnostic accuracy studies and ensuring well-powered and resource-efficient designs.

P49. Evaluation of diagnostic tests with spatially or temporally clustered data, part 2: Scoping review of different methods for estimating diagnostic accuracy for clustered data

Philipp Weber[1], Nicole Rübsamen[2], Julia Böhnke[2], André Karch[2], **Antonia Zapf**[1]
[1]University Medical Center Hamburg-Eppendorf, [2]University of Münster

Background: In diagnostic accuracy studies with temporally or spatially clustered data, dependencies within one individual should be taken into account, depending on the estimand selected. However, the literature shows that in practice the dependencies are generally ignored or the data are aggregated per individual [1–2]. Presumed reasons are that the available statistical methods are difficult to find or difficult to implement for applicants.

Objectives: To identify the different methodological approaches, discuss their respective statistical properties and assess their applicability.

Methods: We conducted a method review in PubMed with corresponding search terms (e.g. “diagnostic accuracy” and “clustered” or “longitudinal”) and added snowballing results. We differentiated between the area under the ROC curve (AUC) on the one hand and sensitivity and specificity as co-primary endpoints on the other.

Results: For the AUC, we found 16 articles (seven methodological approaches). A distinction can be made as to whether the result of the index test is modeled, the ROC curve or the AUC directly [e.g. 3–4]. For sensitivity and specificity, we found 14 articles (six methodological approaches). A distinction can be made between one-model and two-model approaches, in which sensitivity and specificity are modeled either together or separately [5]. In hierarchical models, on the other hand, the two endpoints are first modeled separately and then combined in a bivariate term. In addition, a distinction can be made between parametric, semi-parametric and non-parametric approaches. No approach is absolutely flexible and robust; all have their strengths, but also their limitations. Most of them are implemented in R. We did not find a systematic comparison of the approaches.

Conclusions: There are several methods for appropriately analysing diagnostic studies with clustered data. However, these are rarely used in practice. As no recommendations can currently be made due to the lack of systematic comparisons of the methods, the next step is to perform simulation studies to close this gap.

Funding: German Research Foundation [539658720 "ClusterDiag"]

Remark: This presentation should be given together with “Evaluation of diagnostic tests with spatially or temporally clustered data, part 1”.

References:
[1] doi.org/10.1186/s12968-023-00949-6
[2] doi.org/10.2196/28974
[3] doi.org/10.1093/biostatistics/kxy010
[4] doi.org/10.1177/096228029800700402
[5] doi.org/10.1148/radiol.12120509

P50. Measurement Error: Unlocking Estimates of Test Variability From Routine Data. A Simulation Study of Methods for Statistical Analysis

Simon Baldwin[1,2,3], Alice Sitch[1,2], Susan Mollan[3,4], Balazs Baranyi[3,4], Jonathan J Deeks[1,2]
[1]University of Birmingham, [2]NIHR Birmingham Biomedical Research Centre, [3]University Hospitals Birmingham, [4]INSIGHT Health Data Research Hub For Eye Health

Background: In case-studies (CS) to estimate measurement error from routinely collected biomarker data, differences were identified in the estimates from linear mixed effects (LME) [1], baseline-pairs [2], and autocorrelation [3] models; there were important differences dependent on clinical scenario. In CS1 (blood pressure in children), estimates were similar across the methods. In CS2 (serum albumin in adults with cholangitis), convergence issues affected the autocorrelation model, and estimates increased with disease progression for the LME and baseline-pairs models. In CS3 (patients with stable ocular hypertension), estimates of the measurement error in retinal nerve thickness (RNFL_G, μm^2) were similar for the LME and baseline-pairs models (~8μm^2), but ~2x smaller for the autocorrelation model (~4μm^2). The magnitude of these differences are of concern, and potential biases in the estimates require evaluation.

Objective: Investigate bias in estimates of the measurement error from statistical methods.

Methods: A simulation study was undertaken to assess how estimates from the three methods differ according to data characteristics. The base-scenario considered the results of CS3 (3800 repetitions), varying: number of patients (30-1920); follow-up frequency (3-12 time-points over 6y); and within-individual -to- measurement error variance ratio (1:4 to 48:4). Measurement error estimates were generated for each model and compared.

Results: For the base-scenario, the LME and baseline-pairs models overestimated the measurement error by 85-126%; the percentage bias in the autocorrelation model was between +/-0.24%. Increasing the patient numbers to 1920 decreased the empirical SEs by ~2.9x; decreasing the number to 30 reversed this effect. The autocorrelation model was sensitive to follow-up frequency (3 time-points = 16.7% convergence). There were no convergence issues for the LME and baseline-pairs models with 3 follow-up time-points; the convergence was >99.7% for all three models when fitted to data with ≥4 time-points. Lower ratios of within-individual -to- measurement error variance resulted in truer estimates of the measurement error for the LME and baseline-pairs models, compared to higher ratios.

Conclusions: The autocorrelation approach was least susceptible to bias in estimating measurement error. However, in settings where the within-individual variance will likely be low (relative to measurement error), variability estimates are more similar across the three methods.

P51. Rationale for selecting number of ML algorithms when developing a clinical prediction model: a systematic review

Ram Bajpai[1]
[1]*Keele University*

Background: Machine learning (ML), which is a subset of artificial intelligence and includes a collection of techniques, uses computationally intensive data-driven approaches to develop a predictive algorithm (or model) that can be used for early diagnosis or future outcomes of a health condition. These ML techniques have shown promising abilities to handle complex health data and provide future predictions. However, research showed that these techniques work in very different ways depending on what data is used, its complexity, and the method used to develop the prediction tool. There are thousands of ML algorithms to choose from, and there is no sure way to determine which will be the best for a given dataset. Researchers often train several ML models simultaneously and then compare each other for better fit to select a single model that can be used in practice.

Objectives: The objective of this systematic review is to understand why many ML algorithms are selected by researchers when developing a clinical predictive model.

Methods: This systematic review will be reported according to PRISMA 2020 guidelines. We will include any clinical prediction model study that was developed using ML techniques across all medical domains. A literature search has been conducted for articles published between January 2024 and December 2024 in the PubMed. We selected only one electronic database as a rough PubMed search indicated over 500 ML prediction model studies are published in 2024 and growing exponentially. We will include studies with any study design and data source, all patient-related health outcomes, all outcome formats and restricted to humans only. Information on the rationale for the overall machine learning approach and the selection of the number of ML models will be extracted. Additionally, study level characteristics and prediction model related information will also be extracted. Results will be summarised as percentages (with 95% confidence intervals [CIs]), medians with interquartile range (IQR), including a narrative synthesis.

Conclusions: Findings from this systematic review will provide a better understanding of why many ML algorithms are selected by researchers when developing a prediction model.

P52. Patient and public involvement in prognostic modelling papers - a missed opportunity?

Hannah Cooper[1], Lizzie Fisher[1], Nickson Murunga[1], Louise Haddon[1]
[1]*University Of Leicester*

Background: Patient and public involvement (PPI) in research is defined as “research that is done with; or ‘by’ the public, not ‘to’, ‘for’ or ‘about’ them”. By understanding the lived experience of conditions, research is more meaningful. PPI is well established in applied health and care research but less so within statistical methods research. Prognostic modelling is a form of applied statistical methods research usually conducted by a statistician. It involves developing statistical models to predict future events based on an individual’s characteristics/risk factors, e.g. the 10-year risk of cardiovascular disease. The most recent TRIPOD reporting guidelines, 2024, require authors report PPI conducted in the study or declare no involvement.

Objectives: To explore if PPI is reported within papers describing the development of prognostic models.

Methods: Taking five recent systematic reviews of prognostic development studies in a range of clinical areas (Caesarean births (n=63), Kidney disease (n=42), Gestational diabetes (n=15), Blood transfusions (n=66), Post-surgical pain (n=14)). For each paper included in these systematic reviews, information on PPI was extracted using a form based on the GRIPPS checklist.

Results: Of the 200 papers included in the five reviews, full texts were accessible for 185. Papers were published between 1986 till 2023 with a high frequency of papers included being published from 2016. One paper mentioned PPI, stating no involvement. Therefore, none of the 185 studies reported carrying out any PPI.

Conclusion: Prognostic models are developed for use in clinical practice, to improve patient outcomes, yet none of the studies we assessed reported conducting PPI. This represents a major missed opportunity for ensuring relevance and applicability to those whom the models aim to serve. The results found may be due to the selective nature of the papers reviewed and a full systematic review is warranted. Additionally, these studies may have conducted PPI but failed to report this, hence the recent addition of PPI to the TRIPOD Checklist.

Most papers were mainly published prior to 2024 which may explain the lack of PPI reporting, if this review was repeated in 3 years, we expect papers would be following TRIPOD guidelines and report PPI.

P53. Simulated Application to evaluate models predicting prior risk using randomized controlled trial data

Guiyou Yang[1], Wessel Ganzevoort[2], Sanne Gordijn[3], Thomas Bernardes[4], Gerton Lunter[1], **Henk Groen**[1]

[1]University Medical Centre Groningen, department of Epidemiology, [2]Amsterdam University Medical Center, department of Obstetrics and Gynaecology, [3]University Medical Center Groningen, department of Obstetrics and Gynaecology, [4]Centro de Saúde Córrego Grande, Secretaria Municipal de Saúde de Florianópolis

Background: For any prediction model, it is wise to perform a preliminary evaluation of its clinical impact before undertaking a prospective clinical impact study. Models predicting treatment benefit can be assessed using randomized control trial (RCT) data. However, such models are rare as they typically require extensive data on the effect of treatment. Models predicting risk without treatment (prior risk model, PRM) are more widely available, and are typically assessed using decision curve analysis, which however does not consider the outcome after treatment. We propose a new approach, termed Simulated Application, that uses RCT data for the preliminary evaluation of models predicting prior risk, to assess actual patient outcomes following model-guided treatment.

Methods: We introduce the Simulated Application framework in the context of the two related methods. To illustrate the procedure of Simulated Application, we apply it to a published obstetric care model, with immediate delivery as the treatment and expectant management as the alternative. We provide a matrix of seven treatment benefit thresholds, quantifying deterioration events avoided by treatment, and five prior risk thresholds to demonstrate how the model impact depends on these entities. The impact is summarized in adversity, reflecting negative effects of treatment and adverse outcomes.

Results: We show that Simulated Application allows the adversity of different treatment policies based on a single PRM to be calculated and compared. In the example, the model-guided treatment policy did not result in statistically significantly lower adversity compared to the "treat all" or “treat none” policies simultaneously, i.e. at the same combination of prior risk threshold and treatment benefit threshold, and we conclude that the data in this example did not provide support for a prospective clinical impact study of this model.

Conclusions: When data from RCTs within the scope of a prediction model are available, Simulated Application can be used to assess the impact of implementing the prior risk model for a range of threshold values for treatment benefit and prior risk. The results can be used to guide selective prospective impact studies, at plausible values for treatment benefit and prior risk thresholds.

P54. Methodology challenges of developing prognostic models in rare diseases: insights and benefits from collaborating with patient and public involvement representatives

Laura Kirton[1], Richard Riley[2,3], Piers Gaunt[1], Bernadette Brennan[4], Kym Snell[2,3]
[1]Cancer Research UK Clinical Trials Unit, College of Medicine and Health, University Of Birmingham, [2]Department of Applied Health Sciences, School of Health Sciences, College of Medicine and Health, University of Birmingham, [3]National Institute for Health and Care Research (NIHR) Birmingham Biomedical Research Centre, [4]Royal Manchester Children's Hospital

Background: Conducting clinical research in rare diseases offers practical and methodological challenges, due to the limited available sample size. This is a particular issue for prognostic modelling, with concerns of model overfitting and large uncertainty of predictions. To help mitigate these issues, patient and public involvement (PPI) is valuable.

Objectives: To highlight challenges faced when developing a prognostic model in rare diseases and to describe insights from PPI representatives to statistical methodology decisions.

Methods: A prognostic model was developed in a rare paediatric cancer, Ewing sarcoma, using EE2012 clinical trial data. The model incorporated prognostic factors identified from literature and was used to estimate an individual's risk of event-free survival at clinically important time points. Different methods of model selection and penalisation for model overfitting were considered, using bootstrapping to assess model stability, prediction uncertainty and predictive performance through internal validation. Further, decision curve analyses examined the models' clinical utility at relevant risk thresholds. PPI were consulted during the protocol and study design, model development and deciding on its clinical applicability.

Results: The EE2012 dataset included 640 patients, meaning only 11 predictor parameters for model selection could be included in accordance with sample size guidance; PPI endorsed the chosen predictor parameters. Minimal differences were observed between the different model development approaches in terms of their stability and performance, and there was large uncertainty in the models' predictions. Discussions with PPI and clinicians established acceptable limits of model stability and demonstrated the importance of presenting individualised risk estimates alongside uncertainty intervals. Additionally, they identified relevant risk thresholds for decision making. Overall, the models developed were deemed by PPI and clinicians to be better than having no individualised risk predictions, but further research is warranted to reduce prediction uncertainty and improve clinical utility.

Conclusions: With limited available sample sizes due to rare diseases, it can be difficult to develop a reliable prognostic model with high predictive performance, as it limits the number of predictor parameters and increases model instability and imprecision. However, even with uncertain predictions, PPI groups may still value having individual-level predictive information, rather than focusing on the population risk.

P55. Sequential sample size calculations for developing clinical prediction models: learning curves suggest larger datasets are needed for individual-level stability

Amardeep Legha[1,2], Joie Ensor[1,2], Ben Van Calster[3,4], Evangelia Christodoulou[5], Lucinda Archer[1,2], Rebecca Whittle[1,2], Kym I.E. Snell[1,2], Paula Dhiman[6], Gary Collins[6], Richard D Riley[1,2]

[1]Department of Applied Health Sciences, School of Health Sciences, College of Medicine and Health, University of Birmingham, [2]National Institute for Health and Care Research (NIHR) Birmingham Biomedical Research Centre, [3]Leuven Unit for Health Technology Assessment Research (LUHTAR), KU Leuven, [4]Department of Biomedical Data Sciences, Leiden University Medical Center, [5]German Cancer Research Center (DKFZ) Heidelberg, Division of Intelligent Medical Systems, [6]Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford

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

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

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

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

sample size is strongly dependent on the level of prediction and classification instability deemed acceptable by stakeholders.

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

P56. Prediction Stability of Survival Models

Sara Matijevic[1], Christopher Yau[1]
[1]*University Of Oxford*

Background: In clinical settings, survival prediction models are essential for estimating patient outcomes over time, such as time to disease progression or mortality risk. Advancements in statistical and machine learning methods have expanded the number of available survival models. However, many of these lack the stability required for clinical use, as their outputs heavily depend on the specific development data sample, making predictions highly variable if a different dataset is used. This instability can be detrimental to patient health, as model predictions are crucial for guiding treatment options.

Objective and Methods: In this study, we evaluated the stability of six survival models: Cox Proportional Hazards model, Weibull model, Bayesian Weibull model, Random Survival Forest, DeepSurv, and DeSurv. Using synthetic data and a bootstrapping framework, we first assessed model stability across four levels: consistency of population-level mean predictions, stability in the distribution of predictions, robustness within patient subgroups, and reliability of individual predictions. We then concentrated specifically on the fourth level, individual prediction stability, which is arguably the most clinically relevant. To examine this, we used the SUPPORT dataset and evaluated stability through mean absolute prediction error (MAPE), prediction instability plots, calibration instability plots, and MAPE instability plots.

Results: Our findings indicate that classical statistical models, such as the Cox Proportional Hazards and Weibull models, provide more stable predictions than machine learning and deep learning survival models. Specifically, the average MAPE for the Cox model was 0.0105, significantly lower than DeSurv's average MAPE of 0.0908, illustrating greater variability in the deep learning model's predictions. Plots further corroborated these findings, with the Cox model demonstrating consistently lower prediction and calibration instability. This pattern was also evident in subgroup stability analyses, where the Cox model maintained lower MAPE values across all subgroups.

Conclusion: As the volume of patient data continues to grow, deep learning survival models represent a scalable and efficient approach for deriving novel inferences. More work is needed to improve the prediction stability of deep learning survival models, so they can achieve reliability comparable to their classical statistical counterparts for safe use in clinical settings and better patient outcomes.

P57. Improving Prediction Stability in Deep Learning Models with Bootstrapping Regularisation

Sara Matijevic[1], Christopher Yau[1]
[1]*University Of Oxford*

Background: Deep learning models are increasingly used in clinical settings to estimate patient outcomes, however, many lack the prediction stability necessary for reliable use. Their predictions often vary significantly depending on the development dataset, which can lead to worse patient outcomes and clinicians' mistrust in the model.

Objective and Methods: In this study, we investigated the use of bootstrapping as a regularisation technique to improve the prediction stability of a deep learning model. We evaluated a Stability model against a Simple model on the GUSTO-I (41,021 participants) and Framingham (4,434 participants) datasets. Both models used a deep neural network architecture comprising three fully connected layers with ReLU activations and a sigmoid output for binary classification. Binary cross-entropy loss was used, with the Stability model incorporating an additional regularisation term to minimise deviations between its predictions and the predictions of 100 models trained on bootstrapped data, sampled from a total of 200. The models were assessed using the mean absolute difference (MAD) from the medians of bootstrapped predictions, and by assessing the statistical significance of prediction deviations from the bootstrapped prediction distribution.

Results: On the GUSTO-I dataset, the Stability model had a MAD of 0.019 versus 0.059 for the Simple model, aligning predictions better for 97.54% of participants. On the Framingham dataset, the Stability model achieved a MAD of 0.057 compared to 0.088, with better alignment for 80.69% of participants. For GUSTO-I, the Stability model classified 13.89% of predictions as significantly deviating from the median of bootstrapped predictions, compared to 87.06% for the Simple model. For the Framingham dataset, the figures were 21.35% and 55.04%, respectively. These results are further corroborated by violin plots, which demonstrate the Stability model's closer alignment with bootstrapped prediction distribution peaks compared to the Simple model.

Conclusion: The Stability model was able to improve prediction stability through bootstrapping, especially important in settings when data is more limited as with the Framingham dataset. By regularising predictions to align with bootstrapped distributions, the model achieved greater robustness than the standard training methods. These results highlight the importance of incorporating stability considerations in predictive modelling, particularly in clinical decision-making.

P58. Treatment effect estimation with counterfactual prediction using individual treatment plans: theory and application in radiotherapy

Lotta M. Meijerink[1], Artuur M. Leeuwenberg[1], Jungyeon Choi[1], Johannes A. Langendijk[2], Judith van Loon[3], Remi A. Nout[4], Johannes B. Reitsma[1], Karel G.M. Moons[1], Ewoud Schuit[1]

[1]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, [2]Department of Radiation Oncology, University of Groningen, University Medical Center Groningen, [3]Department of Radiation Oncology (Maastro), Maastricht University Medical Centre, [4]Department of Radiotherapy, Erasmus MC Cancer Institute, University Medical Center Rotterdam

Background: Within the context of radiotherapy, 'model-based clinical evaluation' was proposed, as an alternative approach to assess the benefit of new radiation technologies regarding radiation-induced toxicities, when RCTs are not feasible or ethical. In this approach, the average treatment effect is estimated by comparing the actual observed outcomes under a new treatment technology with predicted outcomes for the same individuals if they had received the comparator treatment. The approach takes advantage of an existing prediction model that incorporates not only patient characteristics as predictors, but also individual treatment details, such as the radiation dose to normal tissues from an individual treatment plan under the comparator treatment technology.

Objectives: To provide methodological guidance to potential users of this model-based clinical evaluation approach, both within and outside of radiotherapy.

Methods: We used principles from causal inference to formalize the approach and systematically identify the conditions necessary for the validity of the estimated average treatment effect. Furthermore, we described sensitivity analyses and strategies to assess the plausibility of the conditions, inspired by literature on prediction model validation and updating.
To illustrate the approach, and to discuss the plausibility of the identifiability conditions, we applied the approach to data of head- and neck cancer patients, to evaluate the potential benefit of proton therapy over photon therapy in reducing radiation-induced toxicity.

Results: Important conditions of the model-based clinical evaluation approach include model quality in the development population, reliability of the individual treatment plan under the comparator treatment technology, and transportability, i.e. that the conditional outcome risks in the development population match those in the target population. We show that model calibration - under some circumstances - can provide evidence of complying with the transportability assumption, and that model recalibration can be used as a type of sensitivity analysis.

Conclusions: The model-based clinical evaluation approach, which uses existing prediction models for causal inference, holds potential in evaluating new personalized treatment strategies or technologies. However, like all non-randomized designs, it has risks and relies on assumptions. Our guidance and illustrative case study provide researchers with a framework to assess the approach's suitability and implement it effectively.

P59. Potential clinical impact of predictive modeling of heterogeneous treatment effects: scoping review of the impact of the PATH Statement

Joe V Selby[1], Carolien CHM Maas[2], Bruce Fireman[3], **David Kent**[4]
[1]Division Of Research, Kaiser Permanente, [2]Erasmus University Medical Center, Rotterdam NL, [3]Division of Research, Kaiser Permanente, Pleasanton CA, [4]Tufts Medical Center

Background: The PATH Statement (2020) proposed predictive modeling for examining heterogeneity in treatment effects (HTE) in randomized clinical trials (RCTs). It distinguished risk modeling, which develops a multivariable model predicting individual baseline risk of study outcomes and examines treatment effects across risk strata, from effect modeling, which directly estimates individual treatment effects from models that include treatment, multiple patient characteristics and interactions of treatment with selected characteristics.

Objectives: To identify, describe and evaluate findings from reports that cite the Statement and present predictive modeling of HTE in RCTs.

Methods: We identified reports using PubMed, Google Scholar, Web of Science, SCOPUS through July 5, 2024. Using double review with adjudication, we assessed consistency with Statement recommendations, credibility of HTE findings (applying criteria adapted from the Instrument to assess Credibility of Effect Modification Analyses (ICEMAN)), and clinical importance of credible findings.

Results: We identified 65 reports (presenting 31 risk models, 41 effect models). Contrary to Statement recommendations, only 25 of 48 studies with positive overall findings included a risk model; most effect models included multiple predictors with little prior evidence for HTE. Claims of HTE were noted in 23 risk modeling and 31 effect modeling reports, but risk modeling met credibility criteria more frequently (87 vs 32 percent). For effect models, external validation of HTE findings was critical in establishing credibility. Credible HTE from either approach was usually judged clinically important (24 of 30). In 19 reports from trials suggesting overall treatment benefits, modeling identified subgroups of 5-67% of patients predicted to experience no benefit or net treatment harm. In five that found no overall benefit, subgroups of 25-60% of patients were nevertheless predicted to benefit.

Conclusions: Multivariable predictive modeling identified credible, clinically important HTE in one third of 65 reports. Risk modeling found credible HTE more frequently; effect modeling analyses were usually exploratory, but external validation served to increase credibility.

Ref: Kent DM, Paulus JK, van Klaveren D, et al. The Predictive Approaches to Treatment effect Heterogeneity (PATH) Statement. Ann Intern Med. 2020;172:35-45.

P60. CHARIOT: A prediction-under-intervention model for cardiovascular primary prevention

Molly Wells[1], Matthew Sperrin[1], Jim Lyness[2], Jennifer Downing[3], David Jenkins[1]
[1]University Of Manchester, [2]Independent patient or public partner, [3]University of Liverpool

Background: Heart failure affects approximately one million individuals across the UK. There is limited guidance on how to monitor heart failure patients but a personalised renal function monitoring tool in people living with heart failure (RENAL-HF) is being developed to facilitate patient monitoring and aid clinical decision making. Longitudinal clinical prediction models are the foundation of the tool and were developed using real-world primary care data. There is growing concern and awareness that such models can underperform in groups of individuals, such as ethnic minorities, raising fairness concerns. This potential unfairness can be a result of biases or other issues in the data, model, or implementation of the model [1].

Objectives: Undertake a scoping review to determine the most appropriate algorithmic fairness definitions, concerns, and unintended consequences for RENAL-HF.

Methods: A scoping review was undertaken to investigate fairness concerns previously described in heart failure prediction and longitudinal modelling literature. Two searches were conducted, one on heart failure and another on longitudinal modelling, using OVID to search databases with restrictions to English language and publications after 2014. The Geersing prediction modelling search terms guided the search, with input from patient and clinical advisors [2]. The longitudinal modelling search was expanded to chronic conditions, guided by advisors, and restricted to longitudinal models using terms from recent reviews.

Results: This review is ongoing, but preliminary data extracted includes fairness definitions, undeserved groups considered, scenarios that can result in health inequalities and any mitigations or solutions proposed for fairness challenges.

Conclusions: This review provides the fairness of current prediction models in heart failure patients and longitudinal modelling literature. Identified concerns will inform development of possible scenarios that may cause unintended consequences when using the RENAL-HF or other decision support tools, ensuring models are appropriate for all patients.

References:
1. Paulus JK, Kent DM. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. NPJ digital medicine. 2020 Jul 30;3(1):99.
2. Geersing GJ, Bouwmeester W, Zuithoff P, Spijker R, Leeflang M, Moons K. Search filters for finding prognostic and diagnostic prediction studies in Medline to enhance systematic reviews. PloS one. 2012 Feb 29;7(2): e32844.

P61. A simulation study investigating the impact of the prediction paradox on clinical prediction model performance

Lorenzo Ficorella[1], Xin Yang[1], Nasim Mavaddat[1], Tim Carver[1], Hend Hassan[1,2], Joe Dennis[1], Jonathan Tyrer[1], CanRisk Investigators[1], Douglas F. Easton[1], Antonis C. Antoniou[1]

[1]University of Cambridge, [2]National Disease Registration Service, National Health Service England

Background: BOADICEA (available via the CanRisk tool, www.canrisk.org) is a widely used algorithm for predicting future breast (BC) and epithelial ovarian (EOC) cancer risks. It incorporates genetic data, family history, mammographic density and lifestyle/hormonal/environmental risk factors. However, it was primarily developed and validated using data from individuals of White ethnicity/European ancestry, which may limit its validity for other ethnicities/ancestries.

Objective: We developed a version of BOADICEA that provides ethnicity-specific risk estimates. We also developed a pipeline to calculate standardised polygenic scores (PGS) for mixed-ancestry individuals.

Methods: We used data from UK Biobank and the Breast Cancer Association Consortium to derive estimates for the distributions of risk factors (lifestyle/hormonal and rare pathogenic variants) in the UK major ethnic groups: White, Black, East Asian, South Asian and Mixed. We also used data from NCRAS (National Cancer Registration and Analysis Service) to extract ethnicity-specific tumour subtype distributions. Two ancestry-specific breast cancer PGS models were developed, based on the published 313 SNP PGS, considering European, African, East Asian, South Asian and Mixed genetic ancestries. We combined these with ethnicity-specific cancer population incidences from NHS England.

Results: The predicted average absolute risks were lower in all the non-White groups than in Whites. The risk distributions, based on the full multifactorial model, were also narrower. Consequently, the proportion of women classified at moderate or high-risk of BC or EOC, according to NICE guidelines, was considerably smaller in non-White women.
For example, among women with unknown cancer family history, the proportion classified as being at moderate or high-risk of developing BC was 18.4% and 1.9% (respectively) in White women but only 4.9% and 0.01% (respectively) in Black women, and only 3.5% and 0.01% (respectively) in Asian women.

Conclusions: The updated version of BOADICEA incorporates estimates that are more appropriate for the UK major ethnic groups. However, validation in prospective studies is necessary, and further efforts will be needed to extend the models to other countries. The observed differences in risk classification among ethnic groups suggest that guidelines for non-White women may need to be revisited.

P62. Incremental prognostic value of blood eosinophils and exhaled nitric oxide (FeNO) to predict asthma attacks: the OxfoRd Asthma attaCk risk scaLE (ORACLE2) patient-level meta-analysis of control arms of 22 randomised trials

Fleur Louise Meulmeester[1], Jacob K Sont[1], Ian D Pavord[2], Ewout W Steyerberg[1,3], Simon Couillard[4]

[1]Dept. of Biomedical Sciences, Leiden University Medical Centre, [2]Nuffield Dept. of Medicine, University of Oxford, [3]Julius Centre, University Medical Centre Utrecht, [4]Université de Sherbrooke

Background: In asthma, type-2 biomarkers blood eosinophil count (BEC) and exhaled nitric oxide (FeNO) identify a higher risk for asthma attacks and anti-inflammatory responsive phenotype. However, prior studies have not demonstrated the combined prognostic values of biomarkers in predicting asthma attacks, likely due to the correlation between biomarkers and limited sample sizes.

Objectives: To assess the relationship of baseline BEC and FeNO and their interaction with future asthma attacks.

Methods: We conducted a systematic review of randomised controlled trials (RCTs) in MEDLINE (1-Jan-1993 to 1-Apr-2021). Individual patient data (IPD) of control arm participants were obtained for meta-analysis from 22 RCTs (n=6,513). Extensive initial data analysis (IDA) identified low missing data rates (0-8.7% per predictor; handled using multiple imputation with mice) and categorisation of some predictors in specific trials, such as age. These categorisations were imputed using post-processing methods to squeeze the imputed value within the category. We investigated the prognostic relationship of BEC and FeNO, and their interaction, with the annualised severe asthma attack rate using negative binomial models and restricted cubic spline (three internal knots). Clustering of observations within studies was accounted for by specifying study as a factor in the fixed-effects model. Analyses were adjusted for asthma severity, Asthma Control Questionnaire-5 (ACQ-5), lung function (FEV1%), attack history in the past year, and follow-up duration as offset variable.

Results: Higher BEC or FeNO values were associated with higher risks of asthma attacks (per 10-fold increase, rate ratio [95% CI]: 1.48 [1.30–1.68] and 1.44 [1.26–1.65], respectively). Observed synergistic effects between BEC and FeNO (P-value for interaction = 0.045) are illustrated by the dissociating spline curves (Figure 3A). High type-2 inflammatory burden was prevalent in the population (Figure 3B).

Conclusions: This large-scale IPD meta-analysis with extensive IDA comprehensively assessed the combined prognostic value of type-2 biomarkers BEC and FeNO in predicting severe asthma attacks. BEC and FeNO are key predictors of severe asthma attacks, with incremental prognostic value. These findings underscore the importance of comprehensive risk stratification in asthma. Future prediction models utilising this IPD dataset may be centred on biomarkers for more individualised clinical decision-making.

PROSPERO: CRD42021245337 Funding: NIHR, QRHRN, FRQS, APQ, SAB, LUF

P63. Predictive Models for Climate-Related emergency health care utilisation: a scoping review

Karina Tapinova[1], Larissa Bernert[1], Andrea Schmidt[2], Felix Durstmüller[2], Verena Fuhrmann[1], Calvin Lukas Kienbacher[1], Dominik Roth[1]
[1]Medical University Of Vienna, [2]Competence Center Climate and Health – Dep. for Climate Resilience and One Health, Austrian National Public Health Institute

BACKGROUND: Emergency healthcare systems require robust predictive models to anticipate and manage resource demands. Climate-related health impacts present a unique modelling challenge due to complex temporal relationships and multiple influencing factors.

OBJECTIVES: To evaluate the methodological approaches used in developing predictive models for climate-related emergency healthcare utilisation.

METHODS: We performed a scoping review of the available literature on scores and tools to predict emergency health care utilisation and clinical outcomes of adult patients in the context of climate and extreme weather conditions. We searched Medline, Ovid, and Embase for publications since 2021 with no language restrictions. We analysed study and population characteristics, meteorological and clinical factors used and effect estimates, and presented results according to the JBI framework of scoping reviews.

RESULTS: Our search identified 1,734 records, 148 of which met inclusion criteria. Data covered 28 different countries, mostly the US (49 studies; 33%), China (34; 23%), and Europe (30; 20%), whereas one study aimed for a worldwide coverage. While most studies focused on high-income regions, there was a notable gap in research from low- and middle-income countries.

Around 10% of studies addressed specific catastrophic events, mainly hurricanes, typhoons and cyclones. Most studies focused on temperature, air pollution, humidity, and precipitation. The most frequent disorders included cardiovascular and respiratory disease. The most common associations were made between temperature and cardiovascular disorders, as well as effects of air pollution on respiratory disorders.

The methodological diversity ranged from simple correlation analyses to sophisticated environmental modelling approaches. Time series approaches, particularly Distributed Lag Non-linear Models, were the most prevalent, followed by Poisson/Quasi-Poisson and logistic regressions. Machine learning techniques such as Random Forests, Neural Networks, and XGBoost were represented less frequently. None of those prediction models were externally validated.

CONCLUSIONS: Our review identifies significant methodological challenges in climate-related healthcare prediction models, particularly regarding external validation and generalizability. We recommend: (1) standardized protocols for external validation across different geographic contexts and (2) improved methods for handling extreme weather events.

TRIAL REGISTRATION: OSF Registry https://doi.org/10.17605/OSF.IO/VER3M

P64. Predicting the 2- and 5-year risk of kidney failure for patients with chronic kidney disease in the United Kingdom: external validation and update of the Kidney Failure Risk Equation

Steven Wambua[1], Mohamed Mhereeg[2], Ayodele Opatola[2], Shamil Haroon[1], Sinead Brophy[2], Richard D Riley[1], Krishnarajah Nirantharakumar[1], Kym I E Snell[1], Nicola Adderley[1], Anthony Fenton[1]
[1]University Of Birmingham, [2]Swansea University

Background: The Kidney Failure Risk Equation (KFRE) is used to estimate the risk of kidney failure in patients with chronic kidney disease (CKD). The NICE CKD guideline recommends using KFRE for risk counselling and to trigger referrals to secondary care. However, the KFRE version in NICE was developed in a single UK region, did not include ethnicity data, used complete case analysis, and lacked external validation.

Objectives: To assess the performance of KFRE, update the model where necessary, and compare the performance of the original and updated versions in external validation.

Methods: We conducted a population-based retrospective cohort study using the CPRD Aurum primary care database. Adults with CKD (two eGFRs <60 ml/min/1.73m^2 at least 3 months apart) and no prior history of kidney failure, were included. We evaluated the performance of KFRE and updated the model by re-estimating the baseline hazard and predictor effects (age, sex, urine ACR, and eGFR) (Model 1), and considered additional predictors (diabetes and ethnicity) (Model 2). Models were developed using cause-specific Cox proportional hazard regression for time to kidney failure, accounting for the competing risk of death. We compared the models at 2 and 5 years in CPRD GOLD and SAIL using measures of discrimination calibration and clinical utility.

Results: Among 1,859,287 CKD patients, 6,299 (0.3%) and 13,546 (0.7%) experienced kidney failure within 2 and 5 years, respectively. Performance measures of KFRE and the updated models in external validation are summarised in Table 1 and the calibration plots are presented in Figure 1. KFRE demonstrated excellent discrimination at 2-and 5-years, but systematically overestimated risk. Updating the baseline hazard and predictor effects (Model 1) improved discrimination at 2- and 5-years. Adding diabetes and ethnicity (Model 2) further improved discrimination. Furthermore, Model 1 demonstrated better calibration, which was improved further in Model 2 in CPRD GOLD, although was slightly reduced in SAIL.

Conclusions: The KFRE version recommended by NICE systematically overestimates the risk of kidney failure in primary care CKD patients. Our updated models, particularly those including diabetes and ethnicity, improve discrimination and calibration. The updated models also demonstrate higher clinical utility than the original KFRE.

P65. Predicting cancer in patients presenting with lung and colorectal symptoms: the Cancer Diagnosis Decision rules (CANDID) study

Sam Wilding[1], Tom Fahey[2], Gareth Griffiths[1], Richard Stevens[3], Clare Bankhead[3], Richard Hobbs[3], Kathleen Potter[4], Irwin Nazareth[5], Danielle Van der Windt[6], Gerry Leydon[7], Sue Broomfield[7], Karen Middleton[7], Alastair Hay[8], Brian Nicholson[3], Lucy Brindle[7], Evangelos Kontopantelis[9], Nadeem Qureshi[10], Tom Marshall[11], Tom Sanders[6], William Hamilton[12], Colin McCowan[13], Frank Sullivan[13], Martin Dawes[14], Brendan Delaney[15], Norbert Donner Banzhott[16], Frank Buntinx[17], Paul Roderick[7], Tony Avery[10], Samantha Hall[18], Paul Little[7]

[1]Cancer Research UK Southampton Clinical Trials Unit, University of Southampton and University Hospitals Southampton NHS Foundation Trust, [2]Royal College of Surgeons, Ireland, [3]University of Oxford, [4]University of Southampton Tissue Bank, [5]University College London, [6]Keele University, [7]University of Southampton, [8]University of Bristol, [9]University of Manchester, [10]University of Nottingham, [11]University of Birmingham, [12]Peninsular Medical School, University of Exeter, [13]University of Dundee, [14]University of British Columbia, [15]Kings College London, [16]Philipps-Universitat Marburg, [17]University of Leuven and Maastricht, [18]PPI representative

Background: Cancer is one of the leading causes of death in England and Wales. Earlier detection of cancer improves quality of life and length of survival, but the signs and symptoms of cancer are harder to identify in the early stages. General Practitioners lack reliable and validated tools for distinguishing which patients have symptoms that require immediate investigation for cancer, and which patients can be reassured that their risk is low. This is reflected in an ever-increasing demand on the urgent suspected cancer referral system, where only 6% of those referred are ultimately diagnosed with a cancer.

Methods: In 2013, the CANDID team ran a DELPHI exercise with 28 participants from primary and secondary care and academic settings across Europe. They identified 50 important items for prediction of lung and colorectal cancers which led to the development of the CANDID study within the NIHR School of Primary Care Research. Participants across England and Wales were recruited to CANDID between 2014 and 2017 if they presented to primary or secondary care with one or more pre-specified lung or colorectal symptoms. Data items identified from the DELPHI exercise were collected, alongside demographic information with optional blood and saliva collected for future research. Participants provided consent for 5-year remote follow-up for cancer diagnoses via cancer registries.

Analysis plan: The study recruited 23,891 participants in England and 536 in Wales. Approval for long-term cancer data has been gained for English participants, and will be sought for Welsh participants in future. Once the cancer data are available, logistic regression will be used to develop multivariable models to predict cancer in the lung and colorectal symptom cohorts respectively. Multiple imputation will be used if appropriate to replace missing values. Fractional polynomials will be used to model

non-linear risks relations with continuous variables. Prediction models will be internally validated to present apparent and optimism-adjusted c-statistics and combinations of sensitivity and specificity.

Conclusion: CANDID is a large-scale study in progress which will develop prediction models for lung and colorectal cancers which have the potential to streamline the pathway between symptom presentation and cancer diagnosis.

P66. Developing and validating single-outcome and multi-outcome prediction models for adverse pregnancy outcomes

Yiran Zhang[1], Glen Martin[2], Tjeerd van Staa[2], Victoria Palin[1]
[1]Division of Developmental Biology and Medicine, University Of Manchester, [2]Division of Informatics, Imaging and Data Science, University Of Manchester

Background: Pregnancy adverse outcomes bring significant risks to the health of both mothers and their babies. Clinical prediction models (CPMs) identify patients at risk by using routinely collected health data. However, the reliability and clinical utility of existing CPMs for pregnancy is limited by poor reporting quality and a lack of external validation. Moreover, some outcomes often occur concurrently or interact, with one outcome potentially increasing the risk of developing others, and clinical decision-making may change depending on the risks of different combinations of outcomes. Most of the existing models focus on single outcomes, failing to provide a holistic view of multiple outcomes in pregnancy, which necessitates the need for addressing multiple outcomes simultaneously to enable timely intervention and prevention.

Objectives: This project seeks to develop new single-outcome CPMs for different pregnancy adverse outcomes, to improve the methodology rigour over existing models. The study then seeks to develop multi-outcome CPMs for simultaneous outcomes using primary care data on different combinations of complications. All models will be internally and externally validated.

Methods: This study will use the primary care Clinical Practice Research Datalink (CPRD), the Pregnancy Registry, and linked hospital episode data for females aged 14-49 between 2000 and 2020. Outcomes include gestational diabetes and hypertension, pre-eclampsia, fetal growth restriction and macrosomia, preterm birth, and stillbirth. Single-outcome CPMs will be developed and validated in line with the TRIPOD-AI guidelines. Development and validation multiple outcome models will use multinomial logistic regression or multi-state models, or other methods determined by the literature.

Results: Preliminary results identified 1.3 million women and 1.6 million pregnancy episodes. 34% of patients had multiple pregnancies. The most common adverse outcome observed was fetal macrosomia, affecting approximately 9.3% of pregnancies. Around 27% of pregnancies had at least one adverse outcome, while 8% experienced two or more adverse outcomes. Development and validation of single-outcome models are ongoing. I will present results on single-outcome models and future plans for developing multi-outcome models.

Conclusions: Addressing gaps and extending function in current CPMs may help enhance comprehensive prediction, support timely interventions, and improve maternal and foetal health outcomes through robust and holistic approaches.

P67. Latent class multivariate probit and latent trait models for evaluating test accuracy without a gold standard: A simulation study

Enzo Cerullo[1], Sean Pinkney[3], Alex Sutton[1,2], Tim Lucas[1], Nicola Cooper[1,2], Hayley Jones[4]

[1]Biostatistics Research Group, Department of Health Sciences, University of Leicester, [2]Complex Reviews Support Unit, University of Leicester & University of Glasgow, [3]Center of Excellence, Omnicom Media Group, [4]Population Health Sciences, Bristol Medical School, University of Bristol

In the context of an imperfect gold standard, latent class modelling can be used to estimate the accuracy of multiple diagnostic or screening tests, as well as disease prevalence. However, the conditional independence assumption is rarely thought to be valid in clinical practice.

More advanced versions of these latent class models can be used to model this conditional dependence, including the latent class multivariate probit (LC-MVP) and latent trait (LC-LT) models, the latter of which has been used much more widely in test accuracy (despite the former being more flexible, as it models the full correlation matrix structure).

To date, no simulation studies directly comparing these two models have been conducted. Hence, we conducted a simulation study comparing the two models. Moreover, we also compared a wide range of prior distributions. We also conducted a subset of simulations forcing the correlations in the LC-MVP model to be positive, by using a recently proposed method (which does not distort the prior - unlike previous methods), allowing us to make a "fairer" comparison to the LC-LT model - which forces these correlations to be positive.

This simulation study will enable us to answer important questions, such as:

- Does one model consistently perform better than the other?
- When does the LT-LC perform better or equal to the LC-MVP? Should one ever use the LT-LC model, given that the more flexible LC-MVP exists?
- Are there good" default" priors to use for the correlation matrices for the LC-MVP? If so, what should they be?
- For the LC-MVP, should we add restrictions on the within-class correlations, such as restricting them to be greater than 0 (like the LC-LT model does)? Or should we allow negative correlations?

Furthermore, since the aforementioned models are unfortunately very computationally demanding and there is a lack of software available to efficiently fit these models, we used our previously developed R package - BayesMVP (https://github.com/CerulloE1996/BayesMVP) - to efficiently conduct this simulation study.
We anticipate that the results of the simulation study will be complete within the next 1-2 months.

P68. Comparing Model-Based Clinical Evaluation Using Counterfactual Prediction to Enriched Randomized Controlled Trials in Evolving Treatment Technologies: A Case Study in Radiotherapy for Cancer

Jungyeon Choi[1], Artuur Leeuwenberg[1], Lotta Meijerink[1], Johannes Langendijk[2], Judith van Loon[3], Remi Nout[4], Johannes Reitsma[1], Karel Moons[1], Ewoud Schuit[1]
[1]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, [2]Department of Radiation Oncology, University Medical Center Groningen, [3]Department of Radiation Oncology (MAASTRO), [4]Department of Radiotherapy, Erasmus Medical Center Cancer Institute

Randomized controlled trials (RCTs) are often considered the gold standard for evaluating the superiority or inferiority of treatments. However, RCTs may not be most suitable in evaluating medical technologies. Medical technologies continue to evolve due to various factors such as learning curves of the medical experts or variability in local working flows. By the time RCT results are available, the evidence may be outdated, failing to reflect the current state of the technology. As an alternative, model-based clinical evaluation (MBCE) has been introduced to assess causal treatment effects of evolving medical technologies. MBCE employs counterfactual prediction, which contrasts observed outcomes under a given treatment with predicted outcomes under alternative treatments.

Newly introduced technologies often provide more targeted care and improved health outcomes but could be less accessible to a broader population. In clinical practice, patients most likely to benefit from these technologies are often selectively chosen. A prominent example is proton-based radiotherapy, which offers more precise tumor control than predominant photon-based techniques, potentially reducing radiation-induced toxicity. In the Netherlands, certain cancer patients are selected for proton therapy over photon therapy based on predicted toxicity risks. Given proton therapy's recent emergence, there is growing interest in estimating its causal benefit by evaluating hypothetical outcomes had patients received photon therapy instead. Although enriched RCTs (where patients with model-based indications for proton therapy are randomized) could be considered, practical challenges make their implementation difficult.

This study compares the validity of enriched RCTs and MBCE in estimating causal treatment effects in the context of evolving radiotherapy technologies. Using simulations based on empirical data, we examine how treatment selection processes influence both methods and assess how model performance impacts the estimation of valid causal treatment effects. Various sample sizes and technological evolution rates will be tested to explore how these factors affect the evaluation process. This study aims to provide insights into whether and to what extent model-based clinical evaluation can complement or substitute for RCTs in the context of evolving medical technologies and to aid reliable and timely decision-making in clinical practice.

P69. PROBAST+AI: An updated quality, risk of bias and applicability assessment tool for prediction models using regression or artificial intelligence methods

Chantelle Cornett[1]
[1]*University Of Manchester*

Background: Prediction models use information about a person to predict their risk of disease. Across health, patients transition between multiple states over time, such as health states or disease progression. Here, multi-state models are crucial, but these models require additional methodological considerations and their application in prediction modelling remains scarce. The methodological state-of-play of these methods in a prediction context has not been summarised.

Objectives: This systematic review aims to summarise and critically evaluate the methodological literature on multi-state models, with a focus on development and validation techniques.

Methods: A comprehensive search strategy was implemented across PubMed, Scopus, Web of Science, arXiv to identify methodological papers on multi-state models up to 7th October 2024. Papers were included if they focused on methodological innovation, such as sample size determination, calibration, or novel computational methods; we excluded purely applied papers. Methodological details were extracted and summarised using thematic analysis.

Results: The search identified 14,788 papers. After the title and abstract screening, there were 448 papers for full-text screening, of which 312 papers were included.

Preliminary findings from these studies reveal the majority of methodological research falls into the following groups:

1. Techniques for estimating transition probabilities, state occupation time, and hazards.
2. Hypothesis testing.
3. Variable selection techniques.

This presentation will overview the themes of methodological work, the limitations/gaps in methodological literature in this space, and outline areas for future work.

Conclusions: Early results highlight progress in the methodological development of multi-state models and emphasise areas requiring further attention, such as more research into sample size and robust validation practices. The final results of this study aim to guide future research and support the adoption of best practices in the use of multi-state models.

P70. Quantifying Between-Study Heterogeneity in Single-Arm Evidence Synthesis

Ulrike Held[1], Lea Bührer[1,2], Beatrix Latal[3], Stefania Iaquinto[1]
[1]Epidemiology, Biostatistics and Prevention Institute, University of Zurich, [2]Centre for Computational Health, Institute of Computational Life Sciences, Zurich University of Applied Sciences (ZHAW), [3]Child Development Center, University Children's Hospital, University of Zurich

In this study we addressed meta-analysis of single-arm observational studies for common outcome types. Between-study heterogeneity is of particular relevance in this situation and is accounted for by estimating the heterogeneity variance parameter τ^2 in a random effects meta-analysis framework. While different estimators for τ^2 have been proposed, there is no clear guideline on which estimator is preferable in specific situations, and systematic evaluations are lacking. In a non-systematic literature review, we assessed which meta-analysis methods are currently used in high-ranked medical journals.

Seven different estimators of the parameter τ^2 were selected to cover common methods used in clinical research and based on availability in the R programming environment. The performance was evaluated regarding mean bias, mean squared error and the proportion of estimates equal to zero. Additionally, we evaluated coverage and bias-eliminated coverage using Wald and Hartung-Knapp confidence intervals. Moreover, prediction intervals were calculated. A neutral comparison simulation study was set up to cover common meta-analysis scenarios for continuous and binary outcomes in a single-arm meta-analysis setting. Additionally, the methods were applied to a case study on infants with congenital heart disease.

By means of simulation, we found imprecision across all heterogeneity variance estimators. The estimation was particularly imprecise in situations where the meta-analysis contained few studies or when addressing binary outcomes with rare events. Most heterogeneity variance estimators produced zero heterogeneity estimates despite the presence of heterogeneity. Interestingly, the estimated overall effects were generally found to be relatively robust to different methods, but prediction intervals varied considerably. Our literature review revealed that statistical literacy of heterogeneity variance estimators in single-arm meta-analyses was low, with over half of the reviewed studies failing to report the chosen estimator.

We conclude that it is not appropriate to rely on a single heterogeneity variance estimator when conducting a single-arm meta-analysis of observational studies. We recommend the use of multiple estimators in a sensitivity analysis, particularly for assessing prediction intervals.

P71. Considering causality in normal tissue complication probability model development: a literature review

Marissa Mulder[1], Jungyeon Choi[1], Lotta Meijerink[1], Wouter van Amsterdam[1], **Artuur Leeuwenberg**[1], Ewoud Schuit[1]
[1]University Medical Center Utrecht, Utrecht University

Background: Normal tissue complication probability (NTCP) models are prognostic prediction models that estimate radiation-induced toxicity in cancer patients based on the radiation dose to healthy tissue. These models can inform patients about their prognosis but are sometimes used for planning the dose distribution to healthy organs or choosing the type of radiation treatment. If NTCP models are to be used to make these clinical decisions, they should accurately capture the causal relation between radiation dose and radiation-induced complication risk to prevent over- or underestimation of predicted complication probabilities which can lead to over- or undertreatment.

Objective: To evaluate the alignment of study aims, potential use, methodology, and causal claims made regarding the association between radiation dose and radiation-induced complications in existing NTCP model development studies in patients with head and neck cancer (HNC).

Methods: We included NTCP model development studies identified in a recent Cochrane review on NTCP models in HNC. Data were extracted on stated model aim, claims for potential model use, adjustments for confounding (when applicable), and use of language implying causality.

Results: Out of 98 evaluated studies, 49 (50.0%) made claims that the models could be used to guide decision making about the radiation treatment (causal claims). However, in only 11 studies (11.2%) this was the initial objective. In total, 36 studies (36.7%) made causal claims in a way that did not match the stated non-causal aim of their research. None of the studies that made causal claims explicitly addressed confounding.

Conclusion: There was frequent misalignment between the stated aim of NTCP model development studies and the claims regarding causality and potential usage of the model. This can lead to insufficient consideration of confounding in the statistical analysis and may lead to over- or undertreatment when the models are used to make treatment decisions. Researchers should precisely express the anticipated use of the NTCP model, and if this requires estimating causal effects, they should consider, discuss, and account for the required assumptions, including confounding.

P72. Comparison of different modelling strategies to enhance equity in post-HCT survival predictions: a case study with CIBMTR data

Xueli Zhang[1], Weibo Kong[1], Zhenhua Wang[1], **Junfeng Wang**[1]
[1]*Shanghai Medical Insight Technology Co.,Ltd*

Background: Clinical prediction models are usually developed for the whole patient population, but fail to account for disparities among patient subgroups, such as socioeconomic status, race, and geography. It is important to develop models that are both precise and fair for patients across diverse subgroups. This study is inspired by the Kaggle competition on equity in predictions hosted by CIBMTR.

Objectives: To compare 3 modelling strategies to develop prediction models for allogeneic HCT, that enhance both accuracy and fairness in model predictions for all race groups.

Methods: The model development dataset (n = 6419) was constructed by stratified random sampling from the original synthetic data (N = 28800), reflecting the proportion of each race group according to the US census.

Three modelling strategies were considered: (1) to develop one single model; (2) to develop separate models for race groups; (2) to develop one single model with inverse probability weighting.
The performance measure to compare the strategies was the stratified C-index (calculated as the mean minus the standard deviation of C-indices among subgroups), which was requested by the Kaggle competition. The original synthetic data was used for model performance evaluation.

Results: The C-index of each race group and the stratified C-index for all 3 strategies are shown in Figure 1.
The Strategy 1 had the best performance (stratified C-index=0.637), while the Strategy 2 had the worst one (stratified C-index=0.595). The probability weighting approach (Strategy 3) had slightly lower performance (stratified C-index=0.624) than Strategy 1. Only in the largest race group (race=White), developing race specific models gained better performance (C-index=0.639) than the single model (C-index= 0.636).

Conclusions: Although it was usually believed that "one size fits all" in prediction models often falls short when applied to diverse subgroups, this case study showed difference results. Developing separate models for subgroups seems not working well, especially for the minorities. More strategies need to be proposed and investigated to further improve the equity in predictions for patients subgroups.

P73. Explainable AI in Healthcare: to Explain, to Predict orto Describe?

Alex Carriero[1], Anne de Hond[1], Bram Cappers[2], Fernando Paulovich[2], Sanne Abeln[3], Karel GM Moons[1], Maarten van Smeden[1]
[1]University Medical Center Utrecht, [2]Eindhoven University of Technology, [3]Utrecht University

Explainable AI methods are designed to provide information about how AI-based models make predictions. In healthcare, there is a widespread expectation that these methods will provide relevant and accurate information about a model's inner-workings to different stakeholders (ranging from patients and healthcare providers to AI and medical guideline developers). This is a challenging endeavor since what qualifies as relevant information may differ greatly depending on the stakeholder. For many stakeholders, relevant explanations are causal in nature, yet, explainable AI methods are typically not well-equipped to deliver such information. Using the well-known framework (Describe, Predict, Explain) and an illustrative example we argue that Explainable AI methods are typically descriptive tools, as they may help to describe how a model works but are limited in their ability to explain why a model works in terms of true underlying biological mechanisms and cause-and-effect relations. This limits the suitability of explainable AI methods to provide actionable advice to patients or to judge the face validity of AI-based models.

P74. Diagnostic accuracy of tests for SARS-CoV-2 acute infection: Distinguishing measurands from target conditions

Anne De Hond[1], Ruurd Kuiper[1], Isa Spiero[1], Yu-Wen Chen[1], Demy Idema[1], Anneke Damen[1], Maarten van Smeden[1], Carl Moons[1], Tuur Leeuwenberg[1]
[1]UMC Utrecht

BACKGROUND: Large Language Models (LLMs) offer promising advancements in healthcare administration and delivery. Thorough validation of these models is crucial for their safe and effective use, as incomplete or incorrect LLM outputs (e.g., hallucinations) can harm patient care. However, the diversity in natural language tasks and outputs complicates the establishment of uniform validation metrics and methods.

OBJECTIVE: This scoping review aims to provide an overview of validation methods for clinical LLMs to facilitate rigorous validation practices and identify gaps in the literature.

METHODS: A systematic search in PubMed and Scopus was conducted to identify relevant LLM validation methods. These methods were categorized across three main (clinical) natural language tasks: summarization (e.g., drafting discharge letters), question answering (e.g., responding to patient inquiries), and chatbots (e.g., discussing upcoming appointments).

RESULTS: From 3534 unique hits, 225 relevant papers were identified. Of these, 67% focused on automatic validation methods (of which 18% on LLM-based validation), 26% on human validation (using human raters or evaluators), and 7% on a mix of the above or other methods. Validation methods for summarization were discussed in 33% of the papers, general methodologies in 27%, chatbot methods in 19%, question answering in 10%, and other tasks like information extraction in 12%. Key validation outcomes included lexical overlap, semantic similarity and faithfulness for summarization; accuracy and response quality for question answering; and overall quality and correctness for chatbots.

CONCLUSIONS: Validation methods, outcomes, and metrics vary significantly across different natural language tasks. This overview lays the groundwork for rigorous validation practices. Further research is needed to identify the most effective validation methods for specific clinical LLM applications and to assess their clinical impact.

P75. Clinical impact assessment of diagnostic and prognostic models: an AI assisted scoping review

Andriy Melnyk[1], Kevin Jenniskens[1], Miriam van der Meulen[1], Kim van der Braak[1], Karel G. Moons[1]

[1]Department of Epidemiology, Julius Center for Health Sciences and Primary Care, UMC Utrecht, Utrecht University

Background: Many diagnostic and prognostic (AI/ML) prediction models are developed every year. These are at the center of informing clinical decision making. After model development and validation, assessment of impact on clinical outcomes through an (empirical) study is an essential step towards implementation in daily practice. It is currently unknown how many empirical clinical impact studies on prediction models have been performed and what their characteristics are.

Objective: To identify and summarize characteristics and quality of empirical clinical impact studies on prediction models across clinical domains.

Methods: PRISMA-ScR guideline was followed for reporting of this review. Medline was searched in December 2024 to identify english-written, empirical studies (e.g., Impact Trials/RCTs, quasi-experimental designs), assessing impact of prediction models on clinical outcomes. An AI assisted screening tool (ASReview) was used to prioritize articles during title and abstract screening. Manual search and screening of 75 articles was performed to provide prior knowledge to the algorithm. BUSCAR methodology was used as stopping rule with a target recall of 95% ($p<0.05$). Data on study design, clinical domain, study participants, impact outcome type, country and whether a statistically significant positive impact was observed were extracted.

Preliminary results: The search yielded 21,394 articles. After an initial screening of 480 articles, 45 (9.4%) were included. Of these, 20 were RCTs and 6 were pre-post studies. 18 RCT protocols were also included to be subsequently snowballed. Most articles were from the USA (n=13), the Netherlands (n=10) and the UK (n=9). The most common clinical domains were cardiology (n=10), infectious disease (n=7) and anesthesiology (n=5). The median number of study participants was 1078 (IQR: 316 – 3908). Impact outcome types were highly heterogeneous, ranging from more clinically relevant (e.g. major adverse cardiac events) to less (e.g. antibiotic prescriptions). In 14/27 studies (52%) the model had a statistically significant positive outcome.

Conclusions: Impact studies prediction models are conducted across clinical domains and differ in their characteristics. This paper provides an overview of impact studies which can be used as a stepping stone for meta-epidemiologic research into this topic.

P76. Evaluation of Semi-Automated Record Screening Methods for Systematic Reviews of Prognosis Studies and Intervention Studies

Isa Spiero[1], A.M. Leeuwenberg[1], Dr. K.G.M. Moons[1], L. Hooft[1,2], J.A.A. Damen[1,2]
[1]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, [2]Cochrane Netherlands, University Medical Center Utrecht, Utrecht University

Background: Systematic reviews (SRs) are used to synthesize evidence through a rigorous, labor-intensive, and costly process. To accelerate the title-abstract screening phase of SRs several artificial intelligence (AI)-based semi-automated screening tools have been developed to reduce workload by prioritizing relevant records. However, their performance is primarily evaluated for SRs of intervention studies, which generally have well-structured abstracts.

Objectives: To evaluate whether screening tool performance is equally effective for SRs of prognosis studies that typically have larger heterogeneity between abstracts compared to SRs of intervention studies.

Methods: We conducted retrospective simulations on 12 previously conducted and published prognosis and intervention reviews, using a semi-automated screening tool. We also evaluated the effects of the review scope (i.e. breadth of the review's research question), number of (relevant) records, and modeling methods (e.g. logistic regression, naive bayes) within the tool. Performance was assessed in terms of recall (i.e. sensitivity), precision at 95% recall (i.e. positive predictive value at 95% recall), and workload reduction (Work Saved over Sampling at 95% recall (WSS@95%)).

Results: The intervention reviews generally reached a recall of 95% earlier in the simulations of the screening process compared to the prognosis reviews (Figure). This was also apparent from the WSS@95% which was (slightly) worse for prognosis reviews (range: 0.324-0.597) than for intervention reviews (range: 0.613-0.895). The precision was higher for prognosis (range: 0.115-0.400) compared to intervention reviews (range: 0.024-0.057). These differences were primarily due to the larger numbers of relevant records in the prognosis reviews which inherently impact the performance metrics. The modeling methods and the scope of the prognosis review did not significantly impact tool performance.

Conclusions: We conclude that the larger abstract heterogeneity of prognosis studies does not substantially affect the effectiveness of (AI-based) semi-automated screening tools for SRs of prognosis studies. Further evaluation studies are needed to enable prospective decisions on reliable use of screening tools to accelerate the conduct of systematic reviews.

P77. Use of Transformer Models for Clinical Prediction – Extending BEHRT for UK Biobank and investigation of prediction performance under different scenarios

Yusuf Yildiz[1], Goran Nenadic[1], Meghna Jani[1], David A. Jenkins[1]
[1]*The University Of Manchester*

Background: Large language models like BEHRT1 have shown potential in modelling Electronic Health Records to predict future instances. These models can represent patient histories by including structured(diagnoses) and unstructured data (doctor notes)2. BEHRT showed superior performance over the state-of-the-art models at the time its developed using a large primary care data set. However, it's unclear if such model and high accuracy can be achieved for hospital data. Developing LLMs requires selecting various decisions and parameters. Parameter choices have been shown to impact model performance, stability and generalisability, but it's unclear the extent this also hold for LLMs. This study aims to implement the BEHRT architecture in the UK Biobank and identify challenges of implementing this model into different dataset. The secondary aim is to assess the impact of parameter choices on prediction performance.

Methods: This study uses UK Biobank data. To capture key features of patient histories, embeddings are created using diagnoses and age at diagnosis. BEHRT workflow included pretraining with masked language modelling(MLM) and fine-tuning for next-diseases prediction across different time frames. Prediction performance was evaluated using Average Precision Score and AUROC. Initially, the original study is replicated using UK Biobank to assess the impact of dataset variability. Subsequently, the model's performance was evaluated to assess the effects of different medical terminologies and data splits.

Results/Conclusion: Our replicated BEHRT model did not achieve as hight predictive performance as performance metrics as the original. Terminologies with bigger vocabularies showed worse performance. Complete separation of the MLM and fine-tuning data resulted with worse performed model. However, most developed models use complete dataset for pre-training and therefore are likely to exhibit overly optimistic performance. A more rigorous and definitive workflow and framework is warranted for LLM development in clinical prediction.
Further work is needed on time-to-event analysis, censoring adjustment, transparent decision-making and computational costs for better integration into clinical prediction.

1. Li, Y. et al. BEHRT: Transformer for Electronic Health Records. Sci. Rep. 10, 7155 (2020).
2. Kraljevic, Z., Yeung, J. A., Bean, D., Teo, J. & Dobson, R. J. Large Language Models for Medical Forecasting -- Foresight 2. Preprint at https://doi.org/10.48550/arXiv.2412.10848 (2024).

P78. How do we evaluate whole genome sequencing for newborn screening for hundreds of conditions?

Karoline Freeman[1], Jacqueline Dinnes[2,3], Corinna Clark[1], Ines Kander[1], Katie Scandrett[2], Dylan Taylor[1], Naila Dracup[1], Rachel Court[1], Furqan Butt[1], Cristina Visintin[4], James R Bonham[5], Graham Shortland[4], David Elliman[8], Anne Mackie[4], Zosia Miedzybrodzka[6], Sian M. Morgan[7], Yemisi Takwoingi[2,3], Bethany Shinkins[1], Aileen Clarke[1], Sian Taylor-Phillips[1]

[1]University of Warwick, [2]University of Birmingham, [3]National Institute for Health and Care Research (NIHR) Birmingham Biomedical Research Centre, [4]UK National Screening Committee, [5]Sheffield Children's NHS Foundation Trust, [6]University of Aberdeen, [7]University Hospital of Wales, [8]Great Ormond Street Hospital

Background: Fifteen genomic projects to sequence a total of 400,000 healthy newborns have been initiated worldwide to evaluate whole genome sequencing (WGS) as a screening tool for hundreds of genetic disorders. No method for synthesising the evidence on the benefits and harms of using WGS to simultaneously screen for hundreds of conditions currently exists.

Objective: To evaluate systematic review approaches to synthesising evidence on the benefits and harms of WGS in newborns for hundreds of genetic conditions for policy advisors.

Methods: We stratified, then randomly sampled 5/200 conditions included in Genomics England's Generation Study, using criteria designed to maximise variability and data availability. We undertook 30 systematic reviews (5x6 review questions) of penetrance (proportion of babies with a specific genotype who develop clinical disease – similar to PPV), detection rate, accuracy, benefit of earlier treatment and benefits and harms of screening (search MEDLINE, Embase, Science Citation Index, Cochrane Library inception to November 2023). Using a non-condition-specific approach, we systematically reviewed genomic studies of newborn screening cohorts reporting penetrance of genetic variants (search inception to January 2024) to identify pathogenic variants with high penetrance for selection of conditions for full evidence review. We assessed the approaches considering review effort and level and quality of evidence.

Results: We screened 19,689 titles and included 268 papers across five conditions addressing the detection rate and treatment benefit in clinical cohorts. No studies were identified reporting penetrance, accuracy, benefits and harms of WGS in screening cohorts. Reviewer time for five conditions was seven months. A team of five reviewers would take over 20 years to conduct similar reviews for 200 conditions. We identified 10 published genomic studies with a total of 76,268 newborns screened. Half (1,449/2,687) of identified genetic cases could not be confirmed by the studies' 'reference standard'. Penetrance data was not available because clinical follow-up was insufficient and genetic findings considered clinically significant were acted on.

Conclusions: Evidence synthesis of benefits and harms of WGS is currently neither feasible nor fruitful. We propose combining evidence collection on penetrance with a staged approach to evaluation, focusing on pathogenic variants with evidence of high penetrance.

P79. Quantifying the Impact of Parkinson's on Walking with Computer Vision

Maedeh Mansoubi[1], **Helen Dawes**[1]
[1]NIHR Exeter BRC, University Of Exeter

Background: Impaired gait and reduced joint range of motion (ROM) are hallmark characteristics of Parkinson's disease (PD), significantly affecting mobility and quality of life. Understanding differences in ROM during walking between individuals with PD and age-matched healthy controls can provide valuable insights into motor impairments and inform therapeutic interventions. This study contributes to the development of a digital clinical outcome assessment (dCOA) for safe mobility and transfers, focusing on the accuracy of ROM as a Technology Reported Outcome (TechRO) during walking and transitions, assessed using a computer vision-based markerless motion-capture system (DigiMotion and DigiGait).

Objectives: To quantify joint ROM and walking parameters in individuals with PD compared to healthy controls using markerless motion capture and evaluate the association of these technology-derived outcomes with clinical measures.

Methods: Ten individuals diagnosed with Parkinson's disease and ten age-matched healthy controls participated in a walking task conducted in a controlled laboratory setting. Video recordings were captured using a 4k_camera positioned at 45 degrees. Joint angles at the hip, knee, and ankle were extracted using DigiMotion Software (V5.01;DigiTherapix), while walking speed and step time were assessed with DigiGait (V1.01). Statistical Parametric Mapping (SPM1D) was employed to analyse joint angle differences across the gait cycle. Walking speed and step time were compared using t-tests. Spearman's rank correlation assessed associations between technology-derived outcomes (SPM results, gait metrics) and Unified Parkinson's Disease Rating Scale (UPDRS) motor scores.

Results: Preliminary findings showed significant differences in joint ROM at the hip, knee, and ankle between individuals with PD and healthy controls during walking. Participants with PD demonstrated reduced ROM, particularly during the swing phase of walking, along with lower walking speeds compared to healthy controls. Significant associations were observed between UPDRS motor scores and technology-derived outcomes.

Conclusion: This study highlights differences in joint ROM and walking speed between individuals with Parkinson's disease and healthy controls. The findings underscore the utility of computer vision-based markerless motion_capture for analysing gait impairments in PD. These tools hold promise for the early detection of motor deficits and the development of targeted rehabilitation strategies to enhance mobility and quality of life in individuals with PD.

P80. MODERNISED Trial design: Cost-effective multi-cancer early detection by measuring patient plasma amino acid cross sections with the Enlighten test.

Thomas Oliver[1], Victoria Goss[1], Sam Wilding[1], Jocelyn Walters[1], Professor Joanne Lord[2], Emma Yates[3], Wesley Sukdao[3], William Herbert[1], Katy McLaughlin[1], Rob Waugh[1], Adam Coleman[4], Irene Soulsby[5], Zaed Hamady[1], Professor Simon Crabb[1], Professor Gareth Griffiths[1], Professor Andy Davies[1]

[1]Southampton Clinical Trials Unit, [2]Southampton Health Technologies Assessments Centre, [3]Proteotype Diagnostics Ltd, [4]ECMC, University of Southampton and University Hospital Southampton NHS Foundation Trust, [5]Patient representative, C/o SCTU, University of Southampton

Cancer remains the leading cause of death in the UK. If found earlier there are more treatment options and patients are more likely to survive 5 years post diagnosis. The Enlighten test has been developed to quickly detect cancers as a measure of amino acid composition in the blood with results returnable within 48 hours from blood draw. Data indicates that the Enlighten test is well placed for improving cancer outcomes as it is particularly sensitive for detecting early-stage cancers.
The NIHR funded MODERNISED study is a prospective, observational, multicenter study that aims to recruit 1350 individuals (1000 cases and 350 controls) with cancer symptoms from 10 solid tumour cancer types (Bladder, Breast, Colorectal, Lung, Melanoma, Oesophageal, Ovarian, Pancreatic, Prostate, Renal). The study will collect a single 4mL blood sample which will be analysed at the Wessex Investigational Sciences Hub (WISH) laboratory.

The primary aim of MODERNISED is to evaluate sensitivity and specificity for the Enlighten test in detecting 10 solid tumour cancer types with deprivation associated gradients in mortality outcomes. Results from this trial will guide the design of a larger, randomised trial to generate evidence for evaluation as a diagnostic test as part of standard of care.

P81. Surrogates For Cancer-Specific Mortality In Cancer Screening Trials: A Systematic Review And Meta-Analysis

Julia Geppert[1], Nefeli Kouppa[2], Adam Brentnall[2], Bethany Shinkins[1], Chris Stinton[1], Karoline Freeman[1], Matthew Randell[1], Samantha Johnson[1], Matejka Rebolj[2], Sian Taylor-Phillips[1]
[1]University of Warwick, [2]Queen Mary University of London

Background: Randomized clinical trials (RCTs) of cancer screening tests typically use cancer-specific mortality as the primary outcome. These trials are expensive and require long follow-up periods. Thus, there is interest in the possibility of using surrogate outcomes instead of cancer-specific mortality to facilitate faster and more efficient trials.

Objectives: Based on previous RCTs, we assessed the association between cancer-specific mortality and six potential surrogate outcomes (late-stage cancer incidence, proportion of late-stage cancers, early-stage cancer incidence, proportion of screen-detected cancers [all or aggressive] and diagnostic yield of screening).

Methods: Using a systematic review approach, we identified all cancer screening trials reporting both mortality and prespecified surrogate outcomes. Association between the surrogates (log relative risk [RR] for the former three, proportion for the latter three) and cancer-specific mortality (log RR) was evaluated using weighted fixed-effects linear model. The 95%CI of the RR of late-stage cancer incidence was investigated as a crude predictor of mortality.

Results: We included 213 articles from 58 trials (62 trial-arm comparisons) covering 10 cancer types. We extracted 1,199 and 1,711 RRs for cancer-specific mortality and cancer stage, respectively, but only 15 data points reported late-stage incidence at a suitable timepoint prior to mortality. For all cancers combined, the correlation between late-stage cancer incidence and cancer-specific mortality was 0.69 (95%CI 0.47-0.84; R^2=0.47; 57 trials). The 95%CI of the late-stage cancer incidence RR included the observed mortality RR in 56/61 (92%) of trial-arm comparisons; this was 14/15 (93%) when only including trials where the surrogate was measured prior to mortality. For the other five evaluated intermediate outcomes, there was strong evidence that they would be poor surrogates for cancer mortality.

Conclusions: This is the largest and most comprehensive meta-analysis of cancer screening surrogate endpoints to date. There is a correlation between late-stage cancer incidence and cancer-specific mortality within and between cancer types and cancer screening tests. Further methodological work is necessary to translate the metrics used for surrogacy in cancer treatment trials to use in the complex intervention of screening. Trials should start collecting data on late-stage incidence and prognostic indices prior to mortality measurement to grow the evidence base.

P82. Assessing Screening, Management and Treatment Efficacy in Patients with Heart Failure and Iron Deficiency in Primary Care

Vijay Maharajan[1], Cynthia Wright Drakesmith[1], Innocent Erone[1], Sarah Haynes[1], Alvin Katumba[1], Joseph Lee[1], Suzanne Maynard[1], Katja Maurer[1], Noemi Roy[1], Margaret Smith[1], Akshay Shah[1], Simon Stanworth[1], Clare Bankhead[1]
[1]*University of Oxford*

Iron deficiency and anaemia are highly prevalent among patients with heart failure (HF) and are associated with poor outcomes, including frequent hospitalisations, reduced quality of life, and increased mortality. The current European Society of Cardiologists (ESC) guidelines recommend regular screening for anaemia and iron deficiency through the assessment of full blood count, serum ferritin concentration, and transferrin saturations in all patients with heart failure. Our primary objective will be assessing adherence to the ESC recommendations in primary care.

The ESC currently recommends intravenous iron replacement. However, recent studies suggest that a lower frequency of administration and a longer period of assessment are required for oral iron to be efficacious. Our secondary objective will be to assess the effectiveness of oral iron in treating iron deficiency in primary care, based on improvement in haematinics.

Objectives:
1. Assess adherence to ESC guidelines for screening and management of iron deficiency in primary care patients with HF.
2. Evaluate the effectiveness of oral iron replacement therapy in improving haematinics, in patients with HF

Methods: A retrospective cohort study using CPRD primary care data from 14/01/2016 to 23/03/2021 will identify adults (>18 years) with incident HF. Outcomes assess the proportion of patients who have met ESC recommendations including initial screening within 1 year, repeated screening and initiation of iron replacement if appropriate. Regression analyses (age, gender, ethnicity, location, and comorbidities) aims to identify potential vulnerable cohorts.

The secondary objective will be addressed using a self-controlled case series to assess treatment efficacy of oral iron replacement. Outcomes include improvement in serum ferritin and haemoglobin (Hb) levels within one year of treatment initiation.

Results: Preliminary data suggest a cohort of 155419 patients with incident HF, with only one third of these patients meeting ESC recommendations for initial screening with Ferritin. We aim to perform multivariate regression including age and co-morbidities to identify at risk cohorts to increase awareness in primary care.

Conclusions: Preliminary findings suggest that adherence to ESC guidelines for screening and treatment of iron deficiency in HF patients is suboptimal. We intend to identify vulnerable cohorts as well as assessing efficacy of oral iron replacement.

P83. Examining the Association between Estimated Prevalence and Diagnostic Test Accuracy Using Directed Acyclic Graphs

Eduard Molins Lleonart[1], Alexandra Jauhiainen
[1]*Astrazeneca*

Background: A precision medicine approach will become increasingly important for personalized rheumatoid arthritis treatment.

Objectives: To evaluate two designs assessing the effect of an investigational drug, as measured by the DAS28-CRP endpoint, versus placebo when a hypothetical predictive marker is present. All participants are dichotomized as marker positive or marker negative before entering the study.

Methods: We compare two methods. Method 1 is based on a marker-stratified design with fixed sample size where enrolled participants are randomly assigned to the investigational or placebo group with equal probability in each marker status. When the trial completes, a subgroup analysis for each marker status is conducted.

Method 2 is a two-stage group sequential design with pre-defined sample size, adapted from published methods. Only marker positive participants are enrolled initially to both treatment arms, avoiding unnecessary exposure to a potentially ineffective treatment for marker negative participants in stage 1. Only if efficacy is shown in the biomarker positive population at an interim analysis, we enroll participants to stage 2, otherwise the study is stopped. At stage 2 we enroll participants from both marker groups into the investigational and placebo arms. A Bayesian approach is used for the probability models. We expect different treatment effects for the investigational drug depending on marker status (predictive biomarker), so the effect probabilities are assessed using conjugate priors. Effects are expected to be homogeneous in the placebo group (i.e. same effect regardless of marker status), so we consider a hierarchical borrowing approach.

Results: Based on simulations, both methods controlled incorrect decision-making below 5% under ineffective true treatment differences between active and placebo. The detection of true effective differences was above 80% using both methods, though Method 1 showed slightly greater posterior probabilities, at expense of including more patients. This since Method 2 tended to stop almost all trials at the first stage in case of futile differences.

Conclusions: Method 2 shows similar power as Method 1 but requires lower sample size on average and can be an attractive design choice for evaluation in rheumatoid arthritis of a drug with a predictive biomarker.
This research was funded by AstraZeneca.

P84. What is the evidence base for claims of accuracy for rapid self-test diagnostics sold in UK retail settings?

Katerina-Vanessa Savva[1], Melody Ni[1], Dimitrios Grammatopoulos[2], George B. Hanna[1], Christopher J. Peters[1]
[1]Imperial College London, [2]Warwick Medical School

Background: Early detection of Alzheimer's disease (AD) remains a critical challenge, with limited translation of biomarkers into clinical practice despite advancements in biomarker discovery (Bernhardt et al., 2023; Schöll et al., 2024). Building on established frameworks and relevant literature, the Biomarker Toolkit was developed and validated to assess biomarkers' clinical potential and guide their progression to clinical use (Bernhardt et al., 2023; Savva et al., 2023). This study uses the Toolkit to bridge the gap between biomarker discovery and AD clinical integration, aiming to identify promising biomarkers and create a roadmap for their development and implementation.

Methods: A comprehensive framework will identify and advance promising biomarkers for early AD detection. The first phase involves a systematic review of blood-based biomarkers, evaluating their relevance and evidence across studies using the PICO framework and Medline/Embase databases. Data analysis will quantify the clinical implementation success rate of identified biomarkers. The Biomarker Toolkit will then score and prioritize biomarkers based on clinical readiness, feasibility, and scalability, addressing challenges at each clinical stage. The iterative nature of the Toolkit ensures flexibility and adaptability, meeting the unique demands of biomarker development. The scoring methodology of the Biomarker Toolkit has been previously described by Savva et al. (2023).

Potential Results and Impact: This study presents a systems-based approach that combines systematic reviews and translational science to create a scalable model for biomarker clinical utilisation. Key outcomes include: i) identifying published blood-based biomarker candidates for early AD detection, ii) quantifying success rates and gaps in biomarker utilisation across the clinical pathway, and iii) identifying the most promising candidates for further development. The Biomarker Toolkit serves as a dynamic pathfinder, prioritising clinically relevant biomarkers and addressing adoption barriers. Applied to real-world biomarkers, it identifies research gaps and provides strategies to guide their clinical trajectory, supporting diagnostic translation, reducing discovery costs, and promoting early diagnosis. Beyond AD, this approach can optimise biomarker discovery and implementation across diseases, improving patient outcomes and benefiting healthcare systems.

P85. The Regulation, use and impact of Direct-to-Consumer Testing in the UK: A Systematic mapping review and mixed methods synthesis

Bircan Ciytak, Clare Davenport, April Coombe, Steven Blackburn, Jon Deeks, Aditya Kale, Finlay Mackenzie, Rachel Marrington, Alex Richter, Jessica Watson
[1]*University Of Birmingham*

Background: Direct-to-Consumer Tests (DTCT) have the potential for positive impacts on individuals and the health economy through increasing autonomy, acceptability, effectiveness and efficiency of testing. However, DTCT have potential to cause harm due to inappropriate testing decisions, errors in test execution and inappropriate post-test decision making. The DTCT market has expanded rapidly since the COVID pandemic and there are concerns that DTCT regulation is not fit for purpose. Mapping of DTCT regulation and guidance and systematic synthesis and assessment of the perspectives of all relevant stakeholders is lacking.

Objectives: To map current UK guidance, policy, and regulation of DTCT.

To synthesise empirical literature, theoretical perspectives and commentaries concerned with the regulation, use and impacts of DTCT in the UK

Methods: Searches were undertaken in November and December 2024 including general medical bibliographic databases, subject specific sources, grey literature, preprints and guidance sources. A wide range of DTCT and Regulation and Guidance terms were developed iteratively into a master search strategy for modification across different sources. Eighteen websites representing professional, regulatory, standard setting, governmental, and non-governmental organisations were also searched. A date limit of 2019 was applied to capture DTCT regulation, guidance and use after the COVID pandemic. Forward and backward citation searches will be undertaken on all included literature.

Included studies will be categorised by literature type and empirical research appraised using tools appropriate to the study design.

Mapping will characterise DTCT regulation and guidance in terms of its content, (pre-market evaluation, marketing, post-market surveillance) quantity, provenance and inter-relationships.

A framework approach will be used to synthesise published research, opinion and commentary using themes including consumer motivation; public trust; useability and accuracy; individual and health system impacts.

Results and Conclusions: This review will identify coherence, inconsistency and gaps in existing regulation and guidance for DTCT. Perspectives about the regulation, use and impact of DTCT will be summarised according to key stakeholder groups (regulators, manufacturers, retailers, consumers and healthcare professionals). Results will be used to inform changes in regulation, guidance and policy relating to DTCT and to identify priorities for future research.

P86. Development, validation and clinical utility: Evaluating methods of reviews and risk prediction tools for pressure injury occurrence (An umbrella review)

Beth Hillier[1,2], Katie Scandrett[1], April Coombe[1,2], Tina Hernandez-Boussard[3], Ewout Steyerberg[4], Yemisi Takwoingi[1,2], Vladica Velickovic[5,6], Jac Dinnes[1,2]

[1]Department of Applied Health Sciences, University Of Birmingham, [2]NIHR Birmingham Biomedical Research Centre, University Hospitals Birmingham NHS Foundation Trust and University of Birmingham, [3]Department of Medicine, Stanford University, [4]Department of Biomedical Data Sciences, Leiden University Medical Center, [5]Evidence Generation Department, HARTMANN GROUP, [6]Institute of Public Health, Medical, Decision Making and Health Technology Assessment, UMIT

Background: Pressure injuries (PIs) place a substantial burden on healthcare systems worldwide. Risk stratification allows preventive interventions to be focused on patients who are at the highest risk.

Objectives: To identify and describe available risk prediction tools for PI occurrence, including the development and validation methods used; and, to evaluate the clinical utility (prognostic accuracy and clinical effectiveness) of such tools.

Methods: Database and website searches were conducted in June 2024 to identify relevant systematic reviews. Studies on paediatric tools, sensor-only tools, or staging/diagnosis of existing PIs were excluded. Methodological quality of reviews was assessed using adapted AMSTAR-2 criteria. Results were described narratively.

Results: Thirty-two systematic reviews were included; seven described the development and validation of risk prediction tools for PI, nineteen reviews assessed prognostic accuracy and eleven assessed clinical effectiveness.

A total of 124 risk prediction tools were identified, developed from as early as 1962 up until 2023. Around half were developed using machine learning (ML) methods. Internal validation methods were not reported within reviews for two-thirds of tools. Only two development/validation reviews included external validations of models. Only one review reported measures of both discrimination and calibration.

Accuracy reviews mostly considered risk prediction tools as ‘tests’ applied at a single point in time at a particular threshold for risk of PI. Although most accuracy reviews set out to evaluate the ‘predictive’ validity of PI risk assessment tools (18/19), 16 of these relied on DTA principles, without any apparent consideration of the time interval between test application and the occurrence of the outcome. Only one accuracy review considered timing of the outcome at all.

Effectiveness reviews mainly focused on reduction of PI incidence. Evaluated tools were predominantly scales derived by clinical experts, as opposed to empirically-derived models

(i.e., with statistical or ML methods). The methodology underlying the development of scales in routine clinical usage is not always explicit.

Conclusions: This overview highlights critical gaps in PI risk prediction tool development and validation, including poor reporting, and limited consideration of timing in prognostic accuracy studies. Further research is needed to establish the clinical effectiveness of appropriately developed and validated tools.

P87. Launching CLEARED (Clinical Evaluation & Assessment for Regulation of Diagnostic tests): An In Vitro Diagnostic (IVD) network Centre for Excellence in Regulatory Science and Innovation (CERSI)

Sue Mallett[1], Mike Messenger[2,3], Robyn Meurant[2], Irfan Hassan[4], Kerrie Davies[3,5], Professor Caroline Moore[1,6], Steve Halligan[1,6], Jane Freeman[3], Jane Medina Haines[4], Alison Smith[3], Shonit Punwani[1,6], Mark Wilcox[3,5]
[1]University College London, [2]Insightful Health, [3]University of Leeds, [4]Psephos Biomedica, [5]Leeds Teaching Hospitals NHS Trust, [6]University College London Hospitals

Background: Efficient and effective regulation of diagnostics not only assists patient and consumer safety, but can improve the nation's health, NHS workload, industrial productivity, as well as supporting regulators, UK innovation, research and economic growth.

Research into the regulation of diagnostic tests is important and time-critical, as the Medicines & Healthcare Products Regulatory Agency (MHRA) is implementing new regulations for Devices and IVDs. This area is challenging due to the diversity of diagnostic test formats (e.g. laboratory, POCT, self-tests, software), functions (diagnosis, screening, prognosis, prediction etc), biochemistry (nucleic acids, proteins, metabolites etc) and benefits (accuracy, costs, time, patient burden). To address this, innovators and manufacturers require support and clear guidance in navigating technical and clinical evidence requirements for regulatory authorisation.

Objectives: Our CLEARED CERSI network brings together academic partners with patients, regulators, industry, and clinicians, to enable more efficient regulatory approval of innovative IVD diagnostic tests in the UK. We will build a collaborative virtual network to enhance regulatory science, enabling UK regulatory IVD authorisations.

Network activities:
Our newly launched CLEARED IVD CERSI network activities include:

- developing network of key IVD regulatory science stakeholders
- network events including training, workshops, surveys and dissemination of innovative approaches, good practice and build expertise
- developing methods for regulatory science and evidence generation
- prioritisation of challenges and needs in co-ordination with MHRA, patients, end-users and Industry.

Team and research priorities: The CLEARED team includes academics and clinicians from University College London, University of Leeds, and Leeds Teaching Hospitals NHS Trust, with regulatory experts from Psephos Ltd and Insightful Health Ltd.

Our core team brings together:

- expertise in regulation of IVD tests (Regulatory Specifications and Requirements), evidence generation and appraisal, including AI-IVD
- statistical and health economic expertise
- clinical expertise and engagement
- existing relationships with key stakeholders, decision makers and implementation partners including MHRA, NICE, industry associations (BIVDA, TOPRA), UKHSA, NHSE NIHR BRCs and HRCs and academic researchers.

We have identified key regulatory barriers where we are exceptionally placed to bring together collaborations to advance regulatory science: good practice for regulatory document development; good practice for post-market surveillance.

P88. Formulation of Delphi-Derived Prioritization Criteria for Transferring Tests from Secondary Care to Primary Care Settings

Natasja Vijfschagt[1], Huibert Burger[1], Marco Blanker[1], Jochen Cals[2], Geert-Jan Geersing[3], Mariska Leeflang[4], Michiel de Boer[1], Gea Holtman[1]
[1]University Medical Center Groningen, [2]Maastricht University, [3]University Medical Center Utrecht, [4]Amsterdam Medical Center

Background: It takes an average of nine years to evaluate new diagnostic tests. Most tests are developed and evaluated in secondary care before they are introduced into primary care. As the evaluation of tests in primary care is both costly and time intensive, it is essential to prioritize their relevance for primary care based on agreed-upon and well-defined criteria. However, a specific list of such criteria has not yet been developed.

Objectives: We aimed to achieve consensus on criteria for prioritizing tests that were validated in secondary care but require further evaluation for use in primary care.

Methods: The process of developing the list with criteria consisted 1) of a literature search and steering group (SG) discussions to identify potential criteria, 2) a Delphi study comprising two rounds with a panel of experts to achieve consensus (≥70% agreement) on the criteria, and 3) final steering group discussions to finalize the set of criteria. The SG consisted of eight Dutch researchers with diverse expertise in diagnostic test evaluation, including general practitioners (GPs) and methodologists. The panel for the Delphi round represented 22 international professionals working in clinical, methodological and policy areas.

Results: The Delphi study started with 32 potential criteria and resulted in a final list of 18 criteria, grouped into three domains: (1) Target Population, evaluating a test's potential for primary care populations including evaluation of e.g. prevalence of a target condition and management-following-testing options; (2) Diagnostic Test of Interest, focusing on the test's role and e.g. potential for ruling in and out and consequences for targeted referrals; and (3) Feasibility and Impact, addressing implementation and broader implications, such as patient and GP acceptability. Before presenting the criteria we included five introductory questions to gather key clinical and epidemiological information about the diagnostic context of the test in primary care.

Conclusions: The expert-based consensus criteria to prioritize tests provide insight into which tests could be prioritized for further evaluation in primary care. This list can be used by researchers in collaboration with GPs. We stress that the prioritization of tests in primary care depends on a complex interplay of criteria.

P89. Finding the right threshold and test combination to diagnose coeliac disease based on serology only: individual participant data review and meta-analysis

Martha Elwenspoek[1], Ruth Brassington[1], Efthymia Derezea[1], Hayley Jones[1], Penny Whiting[1]
[1]*University of Bristol*

Background: Coeliac disease (CD) is an autoimmune disorder, triggered by the protein gluten, found in wheat, rye and barley. Traditionally, CD was diagnosed with a serological test for anti-gluten antibodies followed by confirmatory biopsy. Biopsies are generally safe but are invasive, expensive, and burdensome for patients. Serological tests have a quick turnaround (usually 1-2 weeks), whereas for biopsies there are long wait times.

Objectives: This research aims to determine the accuracy of serological tests at different thresholds, singly and in combination, to optimise alternative diagnostic pathways for coeliac disease diagnosis in children and adults that avoid the need for a biopsy.

Methods: We are conducting an individual participant data (IPD) review and meta-analysis. We have searched Medline, Embase, Web of Science, and ClinicalTrials.gov in August 2020 and November 2024 for cohort studies including adults or children with suspected coeliac disease who have undergone serological tests (index test) and biopsies (reference standard). We have also identified studies through our network of coeliac clinicians. We are contacting authors from eligible studies, requesting serological and biopsy test data, and will form a consortium with consenting authors. We will perform a one-stage IPD bivariate meta-analysis of the accuracy of each test at each threshold, accounting for clustering. PROSPERO: CRD42024582414.

Results: From the 2020 search, we have identified 69 eligible studies and 4 additional studies were identified through our network. Twelve studies have agreed in principle to share data (n=7,086), 16 studies have declined (1 no consent to share data, 1 overlap with other dataset, 14 deleted or no access), and 45 studies have not yet responded. We anticipate a higher success rate for the ca. 30 additional eligible studies published since 2020 identified with the update search.
We will present an update on study recruitment, data collection, and describe the data of studies that are sharing data compared to studies that are not participating.

Conclusions: Our findings will influence future guidelines and practice, potentially making the no-biopsy pathway accessible to more patients. This could mean shorter waiting times for biopsies, quicker diagnosis, and earlier treatment for coeliac disease patients, and cost savings for NHS.

P90. Impact of a diagnostic strategy for appendicitis in children presenting with acute abdominal pain in primary care: design of a cluster randomized trial

Esmee Hogervorst[1], Roderick Venekamp[2], Grietje Knol-de Vries[1], Tamara Platteel[2], Roeland Watjer[3], Nynke Koning[3], Mattijs Numans[3], Guus Blok[1], Eline Naber[1], Michiel de Boer[1], Karin Vermeulen[1], Meefa Hogenes, Eric Hiddink, Chris Tromp, Ramon Gorter[4], Bert Holvast, Jan Hulscher[1], Chantal Everts-Panman, Christine van der Pal, Gera Welker[1], Arne van der Bilt, Drs Elyne Deuring[1], Marjolein Berger[1], Esen Doganer, Huibert Burger[1], Gea Holtman[1]

[1]University Medical Center Groningen, [2]University Medical Center Utrecht, [3]Leiden University Medical Center, [4]Amsterdam University Medical Center

Background: In children with acute abdominal pain, general practitioners (GPs) often face difficulties in distinguishing acute appendicitis (AA) from common self-limiting conditions. This causes 19% of AA cases not being recognized at first GP consultation. Furthermore, about 70% of referred children with acute abdominal pain will not have AA. This results in a significant burden for patients, family and the healthcare system.

Objective: To evaluate the impact of using a diagnostic strategy, consisting of a clinical prediction rule (cPR) including C-reactive point of care testing (CRP-POCT) for AA, among children presenting with acute abdominal pain in primary care, as compared with usual care.

Methods: This is a pragmatic cluster randomized controlled trial performed in primary care with 1:1 permuted-block randomization of general practices. Children aged 4 to 18 years presenting at the GP with acute abdominal pain (≤ 7 days) and without a history of appendectomy, pregnancy, or traumatic cause will be included. GPs in the intervention group will use an externally validated cPR based on 7 symptoms and signs, implemented within the GP information system, followed by CRP-POCT in the medium risk group. GPs in the control group will provide care as usual, i.e. following recommendations of the Dutch College of GPs guideline 'abdominal pain in children', in which no referral criteria specifically for AA are included and CRP-POCT is not recommended. The primary outcome is referral efficiency (proportion of non-referrals among patients with no evidence of AA during 30 days follow-up). Secondary outcomes are safety (proportion of referrals among AA patients during first consultation), proportion of children with CRP-POCT, proportion of children with planned reassessment, child anxiety, parent or child satisfaction, quality of life, and costs. We aim to include a total of 566 children from 150 GP practices to determine an improvement in efficiency of 88% to 95%.

Conclusion: The current study should provide evidence whether, compared to usual care, the diagnostic strategy will decrease the number of non-AA referrals, without delaying a diagnosis of AA. Trial registration: ClinicalTrials.gov: NCT06762275

P91. Lung Surfactant Lipids Quantification using Vibrational Spectroscopy

Zixing (Hings) Luo[1,3], Waseem Ahmed[1], Aneesh Vincent Veluthandath[1], Anthony Postle[3,4], Michael Grocott[2,3,4], Ahilanandan Dushianthan[2,3,4], Ganapathy Senthil Murugan[1]

[1]Optoelectronics Research Centre, University of Southampton, [2]General Intensive Care Unit, University Hospital Southampton NHS Foundation Trust, [3]Perioperative and Critical Care Theme, NIHR Southampton Biomedical Research Centre, University Hospital Southampton NHS Foundation Trust, [4]Clinical and Experimental Sciences, Faculty of Medicine, University of Southampton

Background: Neonatal respiratory distress syndrome (nRDS) is a serious condition that primarily affects preterm infants due to underdeveloped lungs and lack of lung surfactants. However, a timely diagnosis remains a challenge, which delays treatment of nRDS and can result in negative outcomes. Attenuated Total Reflectance Fourier Transform Infrared (ATR-FTIR) spectroscopy combined with machine learning is capable of measuring the lung surfactant concentrations and determining the ratio of two key diagnostic biomarkers - lecithin (L) and sphingomyelin (S) in minutes [1,2].

Objectives: This study aims to predict lipid concentrations in synthetic lipid mixtures consisting of five lipids with partial least squares regression (PLSR) approach from IR spectra collected using a diamond/zinc selenide-based ATR platform that requires a smaller volume of liquid sample.

Methods: Measurements were performed with an Agilent Cary 670 FTIR spectrometer. Scan settings were set to 32 scans at a 4 cm-1 resolution. The mixtures include dipalmitoylphosphatidylcholine (DPPC), 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine (POPC), S, phosphatidylglycerol (PG) and cholesterol (Chol). PLSR was chosen to build the prediction models for surfactant lipids.

Results: An independent test set was used to assess the performance of the PLSR prediction models. Quantification of L (total concentration of DPPC and POPC), S, PG and Chol within the physiological ranges tested was demonstrated (R2 = 0.84, 0.77, 0.70, 0.95, respectively). A prediction interval of L/S ratio of ±0.23 near the diagnostic cut-off region (2.2) for nRDS was generated using the jackknife+-after-bootstrap method (see Supporting Figure 1).

Conclusions: The present study demonstrates the feasibility of ATR-FTIR spectroscopy combined with PLSR to quantify lung maturity biomarkers in liquid form using a small sample volume. This approach can be replicated in a clinical context to predict the concentration of lung maturity biomarkers within minutes to enable rapid diagnosis and timely treatment of nRDS.

References:

[1] Ahmed, W., et al., Towards quantifying biomarkers for respiratory distress in preterm infants: Machine learning on mid infrared spectroscopy of lipid mixtures. Talanta, 2024. 275: p. 126062.
[2] Schousboe, P., et al., Predicting respiratory distress syndrome at birth using fast test based on spectroscopy of gastric aspirates. 1. Biochemical part. Acta Paediatr, 2020. 109(2): p. 280-284.

P92. Modelling studies using trial intermediate outcomes for estimating mortality and morbidity reductions of cancer screening interventions: a methodology review

Vichithranie Madurasinghe[1], Pranshu Mundada[1], Sarah Batson[1], Keith Abrams[1], Bethany Shinkins[1], Sian Taylor-Phillips[1]
[1]*University Of Warwick*

Background: Modelling is sometimes used to combine multiple data sources and predict mortality outcomes of screening interventions. The aim of this study is to review modelling studies using intermediate trial outcome data for estimating clinical benefits and risks of cancer screening interventions.

Methods: Medline (OVID), Embase, Web of Science databases and reference lists of potentially relevant studies identified via electronic searches were searched. A single reviewer screened the titles and abstracts, and two reviewers independently assessed the full text articles and extracted data from articles meeting the study inclusion/exclusion criteria. Disagreements were resolved by consensus. Results were tabulated and presented narratively with descriptive statistics presented where appropriate.

Results:
17 modelling studies were included. 7 (41%) focused on breast cancer screening. Others were models for colorectal (n = 4, 24%), lung (n = 2, 12%), ovarian (n = 1), prostate (n = 1), skin (n = 1) and multiple (n = 1) cancers.

State transition models using Marcov methods (n = 7, 41%) and mathematical formulae-based models applying a set of mathematical formula developed by authors for predicting mortality outcomes (n = 6, 35%) were the most widely used modelling approaches. Fifteen studies (88%) assessed cancer specific mortality outcomes; two (12%) modelled the effects of screening on life-years or quality adjusted life-years. Time horizon ranged from 5 years to lifetime.
Most models (n = 11, 65%,) included in the review were developed post-trial where data extracted from trial reports were put in as the model input parameter(s). The most frequently used intermediate trial outcome was the number of cancers by stage at detection (n = 8, 47%).

Interestingly, all models included in the review predicted a cancer-specific mortality reduction from screening which were not in-line with some trial findings. However most modelling studies (n = 13, 76%) did not provide estimates of uncertainty associated with predicted mortality reductions. Only 6 of 14 studies using new models developed by publication authors described the model validation details.

Conclusion: Methodological drawbacks of models using intermediate trial outcome data for estimating mortality and morbidity outcomes of cancer screening interventions limit the usability of their findings.

P93. Diagnosing dementia with Lewy Bodies: Validation of care pathway and perceived role of cardiac Iodine-123 metaiodobenzylguanidine (cMIBG)

Sara Pretorius[1]
[1]*Nihr Healthtech Research Centre For Diagnostic And Technology Evaluation*

Background: Lewy body dementia is the second most common form of neurodegenerative dementia, associated with a rapid progression and poor prognosis. In the UK, the diagnostic process for dementia with Lewy bodies (DLB) frequently involves a dopaminergic function scan, 123I-FP-CIT SPECT (FP-CIT) otherwise known as DaTSCAN. Another diagnostic tool, 123I-metaiodobenzylguanidine sympathetic innervation scintigraphy; (cMIBG) has been utilised for many years in Japan where clinical studies have demonstrated comparable accuracy for cMIBG compared with DaTSCAN.

Objectives: To understand the current DLB diagnostic pathway, including the current role of DaTSCAN and cMIBG and to gather qualitative data on subjective experience and perspective of DaTSCAN and cMIBG scans including potential use cases for cMIBG.

Methods: A targeted review of clinical guidelines was conducted to determine the diagnostic pathway for patients with DLB and a diagnostic pathway process map was developed. Semi-structured interviews were conducted with clinical experts to validate the proposed care pathway, to elucidate the roles and perceptions of both DaTSCAN and cMIBG in the diagnosis of DLB and to establish use case scenarios for cMIBG.

Results: Interviews validated the DLB diagnostic pathway derived from dementia clinical guidelines. The decision to refer patients for diagnostic scanning was influenced by the clinician's experience and confidence in their clinical judgment, as well as the perceived benefits compared with the potential harms of the diagnostic procedure for the patient. Both DaTSCAN and cMIBG were described as diagnostic facilitators, for confirming DLB diagnosis in patients with suspected DLB. Perceived current and potential uses for cMIBG were: as an adjunct to DaTSCAN where diagnostic uncertainty remains following DatSCAN, for distinguishing DLB from other Parkinsonian syndromes and where DaTSCAN is not appropriate (e.g. for where patient medication precludes use of DaTSCAN).

Conclusions: Diagnosing DLB remains a complex process that relies heavily on clinical expertise, patient histories, and a range of diagnostic tests to enhance certainty. While experienced healthcare professionals may depend on clinical judgment, additional tests play a crucial role in aiding and improving diagnostic accuracy. Cardiac MIBG presents potential applications as both a complement to and an alternative to DaTSCAN in the diagnosis of DLB.

P94. Assessing the value of diagnostic tests: evaluation of a framework for identifying and organising test effects

Jac Dinnes[1], Clare Davenport[1], Bella Harris[1], Lavinia Ferrante di Ruffano[2], Yemisi Takwoingi[1], Sue Mallett[3], Chris Hyde[4], Jon Deeks[1]
[1]University of Birmingham, [2]YHEC, University of York, [3]UCL, [4]University of Exeter

Introduction: Evaluation of the impact of a test requires consideration of the clinical pathway in which the test will be used, identification of important ways in which the test might affect that pathway, and selection of outcomes that adequately assess whether the test's introduction will realise clinical benefit. We aimed to evaluate how well a published test evaluation framework (TEF) captures intended and unintended accuracy and non-accuracy effects of diagnostic technologies typically evaluated in HTAs.

Objectives: To obtain empirical evidence of the frequency and importance of different test effects, including both intended and unintended effects.

Method: A catalogue of HTAs of tests for diagnosis or staging was collated from 7 HTA organisations (published 2010-2020). Evidence reviews underpinning HTAs from two organisations (NICE and MSAC) were purposively sampled. The testing strategies compared were mapped out, potential test effects were identified and categorised using the TEF, and any additional mechanisms that might impact on the test's ability to effect downstream outcomes identified. Extractions were conducted by one reviewer, checked by a second and discussed at roundtable project meetings.

Results: The 45 included reviews reported 50 review questions (denominator for all %s); 76% compared an index and comparator strategy using the same type of diagnostic technology (e.g. IVD versus IVD). The clinical claim for the test was reported in a dedicated section in 56%. All test effects in the claim for the test were mapped to pre-specified outcomes in 46% of reviews. In 82%, additional outcomes were specified that were not included in the clinical claim for the test.

Reviews always considered at least one test effect related to impact (accuracy 98%; therapeutic yield, 84%; treatment effectiveness, 94%). Feasibility and interpretation effects were considered in 80% (either acceptability, 42% and/or test failure rates, 48%). Timing mechanisms were considered by 70% of reviews (time to produce a result, 36%; speed of diagnosis, 46%; time to treatment, 46%).

Conclusions: The TEF provides a useful tool for elicitation of intended and unintended effects of introducing a new test or changing a testing strategy, however HTAs may miss evaluating key aspects of a diagnostic's potential value.

P95. Understanding the difference between risk of bias and concerns regarding applicability in the reference standard domain of QUADAS-2: A methodological review

Jude Holmes[1], Eve Tomlinson[2]
[1]UCL, [2]University of Bristol

Background: Diagnostic test accuracy (DTA) systematic reviews bring together findings from DTA studies to summarise the accuracy of a diagnostic test. Studies included in a DTA review should be assessed for risk of bias and applicability concerns, because a review of biased studies, or studies that do not apply directly to the review question, could result in misleading conclusions. Studies are most commonly assessed with the QUADAS-2 tool. Anecdotal evidence has suggested that researchers sometimes struggle to differentiate between risk of bias and applicability. Here, we investigate this distinction for the reference standard domain.

Objectives: To highlight challenges with the assessment of the applicability of the reference standard within Cochrane DTA reviews, including occurrences in which risk of bias issues are mistaken for applicability issues. We will identify examples of applicability concerns, discuss the distinction between risk of bias and applicability, and explore where the reference standard domain intersects with other domains.

Methods: DTA reviews were eligible for inclusion if they were published in the Cochrane Library, had used QUADAS-2, and had at least one study rated as “high concerns” for applicability of the reference standard domain. From each review, we extracted title, authors, publication date, URL, population, index test(s), target condition, reference standard, outcomes, objective, and the rationale provided by the authors for “high concerns” applicability judgements for the reference standard domain. One reviewer assessed these reasons and coded them as either being correctly or incorrectly related to applicability (i.e. actually a risk of bias issue). We also recorded any other issues that arose as part of the applicability assessment. Review selection and data extraction was checked by a second reviewer.

Results: This review is in progress. We have identified 50 DTA reviews that meet our inclusion criteria. Results will be presented at the conference.

Conclusions: Understanding the use of a risk of bias and applicability tool in research is important to ascertain the usefulness and limitations of the tool. The findings of this review will inform the development of QUADAS-3, which in turn may positively affect the robustness of evidence production.

P97. The trade-off between true positive and false positive test results may depend on the maturity of the diagnostic test

Werner Vach[1]
[1]*Basel Academy For Quality And Research In Medicine*

Several tests are often developed over time to address a specific diagnostic question. It is not unlikely that early tests will mainly diagnose patients who are easy to diagnose. Over time, tests will become better at diagnosing patients who are difficult to diagnose. Difficulty in diagnosis (e.g. due to few symptoms or low signal intensity in imaging) is often associated with greater benefit from available treatment. This means that those patients who are additionally identified by new tests tend to have an increasing benefit over time. The figure illustrates this point.

Consequently, when comparing two tests that are adjacent in time, the same increase in TP results may imply an increasing benefit for the corresponding patients over time. This suggests accepting more FP results over time for the same number of additional TP results. With other words: The trade-off between TP and FP results depends on the maturity of the test.

In my contribution, I discuss some implications of this finding for the design and analysis of comparative accuracy studies. The decision to change the weight (or probability threshold) over time is based on assumptions that should be carefully tested empirically using follow-up information. The possibility that the new TP results may be better considered as test negatives should be discussed conceptually. If decision curves are used to present results, the potential change in probability thresholds over time should be taken into account.

Finally, I will raise the question of whether such a change in weighing over time might be justified even if the benefit does not change over time. The simple wish that difficult-to-diagnose patients have a chance of being diagnosed may justify this.

P98. Effects of Using Natural Language Processing for Cohort Selection from Electronic Health Records on Subsequent Prognostic Prediction Model Performance

Jana Wiesner[1], **Antonia Zapf**[1]
[1]*University Medical Center Hamburg-Eppendorf*

Background: For ethical reasons, sample size planning is essential for confirmatory diagnostic accuracy studies. According to guidelines, sensitivity and specificity are recommended as co-primary endpoints [1]. This means that the type 1 error does not need to be adjusted for multiplicity, but the power of the two endpoints needs to be adjusted to ensure a predefined overall power of the study. There is no software for joint sample size planning.

Objectives: To develop an R shiny app for sample size planning for single-test and comparative diagnostic accuracy studies with sensitivity and specificity as co-primary endpoints.

Methods: The methods for optimal sample size planning for sensitivity and specificity by Stark and Zapf [2] and Stark et al [3] are used. The total power is split between the two endpoints in such a way that the total sample size is as small as possible while ensuring the necessary power, taking into account the prevalence. The R Shiny app has been programmed to be as flexible as possible in terms of study design and can be used by researchers without knowledge of R and without in-depth understanding of the methodology.

Results: The app allows sample size planning for the single-test design and for the paired and unpaired comparative design. In addition, both superiority and non-inferiority hypotheses are possible in the comparative design. It is also possible to calculate the power for a given sample size. There are clear explanations of the parameters to be entered and the results are summarised in a comprehensible way. The app is freely accessible and does not require R to be installed.

Conclusions: The free R-Shiny app allows non-methodologists to optimally plan sample sizes for diagnostic accuracy studies with sensitivity and specificity as co-primary endpoints for a variety of study designs.

1. EMA 2009. Guideline on Clinical Evaluation of Diagnostic Agents.
2. Stark M, Zapf A. DOI: 10.1177/0962280220913588.
3. Stark M et al. DOI: 10.1186/s12874-022-01564-2.

Funding: German Research Foundation [458526380 „EpiAdaptDiag“]

Author Index

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

X

Y

Z

The British In Vitro Diagnostics Association (BIVDA) is the national trade association representing UK-based manufacturers and distributors of in vitro diagnostic (IVD) products.

With a membership of 245 organisations—ranging from global multinationals to micro-enterprises—BIVDA champions the UK IVD sector on both national and international stages. We foster collaboration, influence policy, and enable meaningful partnerships between industry, government, and healthcare providers.

The medical diagnostics sector is one of the most innovative in UK life sciences, contributing over £2 billion to the economy each year and enabling the NHS to carry out 1.5 billion diagnostic tests annually. Our members are at the forefront of essential fields such as genomics, molecular diagnostics, and combatting antimicrobial resistance.

Through strong advocacy and close stakeholder engagement, we ensure diagnostics are placed at the heart of modern, effective healthcare in the UK.

The success of our industry is central to better health outcomes, NHS performance, and ultimately the health and wealth of the UK.

SMART DATA
ANALYSIS AND STATISTICS

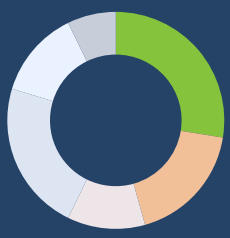
At Smart Data Analysis and Statistics, we specialize in advanced biostatistical analysis and AI-driven solutions for the healthcare and pharmaceutical industries. We provide insights that help optimize clinical trials, improve prediction models, and enhance decision-making. Find out more at **www.fromdatatowisdom.com**

"Thomas Debray is an unmatched mastermind biostatistician... His mastery of advanced statistical methods and his impact in drug industry blend a unique academic knowledge with a visionary and creative gift for business solutions"

Director, Statistical Innovation
Drug Sponsor

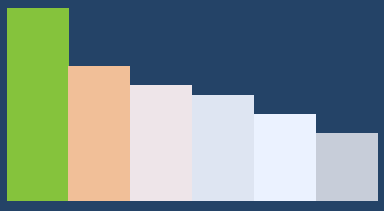
Don't miss out on this unique experience! **Visit us at the booth** for a chance to learn more about cutting-edge AI technology and get a surprise gift!

Our Expertise and Key Contributions



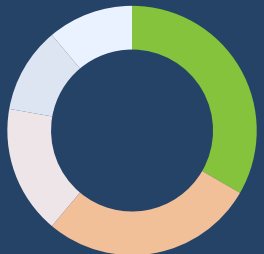
- Meta-analysis and evidence synthesis (38)
- Risk prediction (25)
- Clinical applications (16)
- Causal Inference (31)
- Methodological Guidelines (18)
- Other (10)

How We Deliver Client Succes



- Custom Analytical Tools and Solutions (20)
- Scientific Manuscripts and Research Reports (14)
- Study Planning and Protocol Development (12)
- Quality Assurance (11)
- Project Feasibility and Risk Analysis (9)
- Tailored Biostatistics Training and Workshops (7)

Industries We Serve



- Specialized Service Provider (33.3%)
- Pharmaceutical Company (27.8%)
- Public Health & Policy (16.7%)
- Contract Research Organization (11.1%)
- Biotech (11.1%)