Mini-symposium of the STRATOS initiative at the ISCB/GMDS 2026 conference

Title:

Application of statistical methods needs to improve – on the critical role of guidance for analysis and knowledge translation

Organizers: Anne-Laure Boulesteix (Ludwig-Maximilians-Universität, Munich, Germany)**,** Georg Heinze (Medical University of Vienna, Vienna, Austria), Willi Sauerbrei (Medical Center - University of Freiburg, Freiburg, Germany)

This time, STRATOS is organizing a mini-symposium for two different target groups.

The morning session (9:00–12:30, **Methodological aspects**) is designed in the usual format of a half-day meeting with several shorter talks on behalf of TGs and panels.

In the afternoon (1:30–5:00 PM, **Practical aspects of medical research**), we offer, for the first time, a workshop for clinicians and data analysts with less training and experience in statistical topics. Gary Collins (University of Birmingham, UK) and Matthias Briel (University Hospital Basel, Switzerland) will give keynote presentations.

Program

1 October, 2026

Session 1: 9:00-10:30 and 11:00–12:30: **Methodological Aspects**

Session 1A: 9-00-10:30

Chair: Willi Sauerbrei (Freiburg, Germany)

9:00-9:10: Willi Sauerbrei (Freiburg, Germany): **Introduction**

9:10-9.35: James R. Carpenter (London, UK): **P-values and hypothesis testing: beyond polemics to practical solutions**

9:35-10:00: Michal Abrahamowicz (Montreal, Canada): **How important are the Hazards of Hazard Ratios?**

10:00-10:25: Anne C. M. Thiébaut (Villejuif, France): **Methods for Adjusting for Covariate Measurement Error in Flexible Modelling of Functional Form: Results of a Blinded, Controlled Neutral Comparison Simulation Study**

Session 1B: 11:00–12:30

Chair: Georg Heinze (Vienna, Austria)

11:00-11:25: Marianne Huebner (Michigan, USA): **Statistical analysis plan with initial data analysis (SAPI): Validating the SAPI checklist in analysis projects**

11:25-11:50: Doranne Thomassen (Leiden, the Netherlands): **Comparing five frameworks that may be used to define estimands: friends not foes**

11:50-12:15: Els Goetghebeur (Ghent, Belgium): **Sensitivity analyses for missing data in observational studies: A practical guide for planning, conducting and reporting under MAR and beyond**

12:15-12:30: **Discussion 1** (**Chair:** Georg Heinze)

Session 2: 13:30 – 15:00 and 15.30–17:00: **Practical Aspects of Medical Research**

Session 2A: 13:30 – 15:00

Chair: Willi Sauerbrei

13:30-13:40: Willi Sauerbrei: **Introduction**

13:40-14:10: Gary Collins (University of Birmingham, UK): **The importance of methodology and transparency in predictive AI**

14:10-14:27: Riccardo De Bin (Oslo, Norway): **From traditional statistical modelling to machine learning algorithms: a perspective from the STRATOS TG9 high-dimensional data viewpoint**

14:27-14:44: Ben van Calster (Leuven, Belgium): **Performance evaluation of predictive AI models to support medical decisions: overview and guidance**

14:44-15:01: Peggy Sekula (Freiburg, Germany): **Designing Prognostic Factor Studies: Key Concepts and Practical Guidance**

Session 2B: 15.30–17:00

Chair: Anne-Laure Boulesteix (Munich, Germany)

15:30-16:00: Matthias Briel: (University Hospital Basel, Switzerland): **Meta-research to**

improve evidence generation and synthesis

16:00-16:15: Theresa Ullmann (Vienna, Austria): **The Problem with Univariable Selection in Regression Modelling — and What to Do Instead**

16:15-16:30: Saskia le Cessie (Leiden, the Netherlands): **Guidelines to the design, analysis and interpretation of patient-reported outcomes in cancer clinical trials**

16:30-16:45: Willi Sauerbrei: **An overview and categorization of papers published in statistical series of medical journals**

16:45-17:00: **Discussion 2** (**Chair:** Anne-Laure Boulesteix)

Abstracts

Session 1A

P-values and hypothesis testing: beyond polemics to practical solutions

James Carpenter (London School of Hygiene & Tropical Medicine, London, UK; University College London, London, UK) on behalf of the p-value working group of the STRATOS initiative.

Despite recent articles highlighting concerns about the ubiquitous use and misleading interpretation of p-values, a glance at the observational research literature confirms problems persist. This prompted the STRATOS initiative to explore practical proposals to improve practice.

Building on recent literature, we distinguish three broad goals for observational data: building descriptive models (e.g. to describe associations and develop hypotheses), building prediction models and addressing causal questions.

For each goal, we discuss study planning considerations, and the key tasks p-values can be used for (e.g. evaluating hypotheses; as tuning parameter in model selection) using examples to highlight key considerations for good practice. These include study registration, thorough initial data analysis, an increased emphasis on reporting of the methodological steps, as well as structured reporting of the results. We argue all are necessary for reproducible research.

How important are the Hazards of Hazard Ratios?

Michal Abrahamowicz (McGill University, Montreal, Canada), Marie-Eve Beauchamp (McGill University, Montral, Canada), Emily Roberts (University of Iowa, USA), and Jeremy Taylor (University of Michigan, USA) for STRATOS Topic Group 8 ‘Survival Analysis’)

In time-to-event data analyses, the hazard ratio (HR) is a frequently used metric of the strength

of the association. In a highly cited commentary, M. Hernán criticized using of Cox proportional hazards (PH) model-based HRs by emphasizing that they suffer from the 'built-in selection bias [*Epidemiology* 2010]. To illustrate this bias, he presented the results of a single real-world randomized trial and argued that the estimated decreases in HR with increasing follow-up time are likely simply due to a latent frailty and, thus, should not be interpreted as the evidence of the true (causal) treatment effect diminishing over time. This article led to major concerns about the validity of Cox PH regression analyses. We rely on comprehensive simulations and real-world analyses to revisit this controversial issue First, we demonstrate that the built-in selection bias is usually quite modest, unless the unmeasured frailty has a very strong impact on survival. Then, in dedicated simulations that closely mimic the data structure of the trial discussed by Hernán, we demonstrate that the observed time-dependent changes in treatment HR were practically impossible to reflect just an unmeasured risk factor, even with an extremely strong effect. This finding calls for alternative explanations, that likely include a combination of decreasing treatment adherence and biological changes associated with prolonged treatment. Overall, our simulation results provide strong verifiable evidence that the concerns about the built-in selection bias, and about modeling and interpretation of HRs, are largely overstated.

Methods for Adjusting for Covariate Measurement Error in Flexible Modelling of Functional Form: Results of a Blinded, Controlled Neutral Comparison Simulation Study

Aris Perperoglou (GlaxoSmithKline, London, UK), Mohammed Sedki (Université Paris-Saclay, UVSQ, Inserm, Gustave Roussy, Villejuif, France), **Anne C.M. Thiébaut** (Université Paris-Saclay, UVSQ, Inserm, Villejuif, France), Steve Ferreira Guerra (McGill University, Montreal, Canada), Paul Gustafson (University of British Columbia, Vancouver, Canada), Frank E. Harrell, Jr. (Vanderbilt University School of Medicine, Nashville, Tennessee, USA), Willi Sauerbrei (University of Freiburg, Germany), Michal Abrahamowicz (McGill University, Montreal, Canada), Laurence S. Freedman (Information Management Services Inc., Calverton, Maryland, USA), on behalf of Selection of variables and functional forms in multivariable analysis (TG2) and Measurement Error and Misclassification Topic Group (TG4) of the STRATOS initiative

Covariates in medical research are often measured with error, biasing estimates of exposure-outcome relationships, especially when these relationships are non-linear. This study compares methods for measurement error correction combined with flexible regression modelling techniques in such non-linear settings.

This blinded, controlled neutral comparison, multi-stage simulation project, a collaboration within the STRATOS initiative (Topic Groups 2 and 4), involved a Data Generation and Evaluation team and three Methods teams. These teams applied Bayesian methods, Multiple

Imputation/Regression Calibration (MI/RC), and Simulation Extrapolation (SIMEX), combined with flexible modelling techniques (B-splines (BS), P-splines (PS), Fractional Polynomials (FP), and Natural Splines (NS)). Datasets featured a binary outcome, a continuous covariate with classical error, and a replicate substudy. The true non-linear functional form, covariate distribution, error variance, and error distribution were initially withheld. Stage 1 used 5 pilot datasets. Stage 2 expanded to 150 unique datasets by varying sample sizes, measurement error (ME) variance, error distribution (Normal, shifted-Gamma), and true functional forms. Stage 3 consisted of simulating independent replications of each combination to quantify the sampling variance of the estimates. Performance was assessed by log Mean Squared Error over pre-selected values of the covariate distribution.

SIMEX methods consistently demonstrated the highest accuracy. P-splines, FPs, and NS generally outperformed BS, especially with SIMEX or Bayesian approaches. Following SIMEX, Bayesian methods (excluding BS) performed best, then RC (excluding BS), and MI. Bayesian BS combinations typically performed poorest, particularly with smaller samples. Accuracy generally improved with larger sample sizes and smaller ME. Linear relationships were estimated most accurately; J-shaped forms were most challenging. Notably, SIMEX was less sensitive to increased ME magnitude and, unlike MI and Bayes, showed no substantial accuracy improvement with larger replication substudy sizes.

This joint work emphasizes the relevance of neutral comparison studies to fairly evaluate statistical methods aimed at addressing a complex analytical challenge, and demonstrates their feasibility through a large collaborative project.

Session 1B

Statistical analysis plan with initial data analysis (SAPI): Validating the SAPI checklist in analysis projects

Marianne Huebner (Michigan State University, USA), Carsten O Schmidt (University of Greifswald, Germany), Lara Lusa (University of Primorska, Slovenia), Georg Heinze ((Medical University of Vienna, Austria)

The SAPI statement (Statistical Analysis Plan for observational studies with Initial data analysis) provides recommendations for developing analysis plans for observational studies. It promotes a systematic process to improve completeness and transparency, and ultimately to better interpretability of research results for evidence-based health decision making. It can be used for

descriptive, predictive, or causal research objectives. The SAPI checklist consists of 27 items comprising administrative information, project background, design and data, variables, main data analysis (MDA), initial data analysis (IDA), evaluation and updates, and supplement with plans for disseminating and archiving research outputs.

We demonstrate these items in several examples of analysis plans and their updates following IDA including investigating functional forms in regression models or estimating learning curves.

In each of these examples, updates of the SAP were implemented after IDA. These examples serve to demonstrate that a full pre-specification of an SAP is rarely possible without data examinations. The SAPI checklist and IDA framework guarantee that these looks are performed while strictly ignoring outcome-predictor associations and that their impact on the final analysis specification is transparently reported.

Comparing five frameworks that may be used to define estimands: friends not foes

Nicholas Bakewell (University of Toronto, Canada), **Doranne Thomassen** (Leiden University Medical Center, the Netherlands), Jonathan Bartlett (London School of Hygiene and Tropical Medicine, UK), Suzanne M Cadarette (University of Toronto, Canada), Saskia le Cessie (Leiden University Medical Center, the Netherlands), Nan van Geloven (Leiden University Medical Center, the Netherlands)

There is a surge of interest in defining estimands, which translate a research question into a precise target quantity for estimation. The methodological literature is however divided over how best to define estimands, which may confuse researchers. To provide practical guidance, we offer a comparative overview of five key frameworks that may be used to define estimands: PICOT (Population, Intervention, Comparator, Outcome, Time), the ICH E9(R1) addendum on estimands, the Target Trial Emulation Framework, the Causal Roadmap, and the STRATOS (STRengthening Analytical Thinking for Observational Studies) Causal Inference Topic Group framework. After outlining each framework, we demonstrate their similarities and divergence by applying them to a common research question. We then highlight key strengths and limitations of each framework for defining estimands. We find that each framework has unique strengths and limitations, reflecting its original purpose and scope. We advocate for an approach in which researchers thoughtfully choose and sometimes combine elements of frameworks to leverage their strengths. Using estimands in this way promotes greater clarity, minimizes ambiguity, and enhances the transparency and interpretability of study results.

This presents joint work of members from STRATOS Topic Groups 7 (Causal Inference), 5 (Study design), and 4 (Measurement error and misclassification): Kelly Van Lancker (Ghent University, Belgium), Els Goetghebeur (Ghent University, Belgium), Pamela Shaw (University of Washington, USA), Emmanuelle Boutmy (Merck HealthCare KGaA, Germany), Tim Morris (Novartis Pharma UK Ltd, UK), Rima Izem (Novartis Pharma AG, Switzerland), Susan Halabi (Duke University, USA), Daniala L Weir (Utrecht University, the Netherlands), Marc Vandemeulebroecke (Bayer BCC AG, Switzerland)

Sensitivity analyses for missing data in observational studies: A practical guide for planning, conducting and reporting under MAR and beyond

Dries Reynders (Ghent University, Ghent, Belgium), Rheanna M. Mainzer (Murdoch Children's Research Institute, Melbourne, Australia), Saskia le Cessie (Leiden University Medical Center, The Netherlands), James R. Carpenter (MRC Clinical Trials Unit, London, UK), **Els Goetghebeur** (Ghent University, Ghent, Belgium), Katherine J. Lee (Murdoch Children's Research Institute, Melbourne, Australia), on behalf of TG1 and TG7 of the STRATOS initiative

In observational studies as in randomized trials with longitudinal outcomes, missing data are hard to avoid. Clear assumptions on the missing data mechanisms are then needed to allow for unbiased estimation of most target estimands. As these assumptions are typically untestable, sensitivity analyses come highly recommended.

Allowing for missingness not at random (MNAR) can be complicated and time-consuming. It may furthermore arrive at bounds on the estimated estimand that are unduly wide. How to usefully accomplish a sensitivity analysis depends on the study design and target estimand. Little guidance is available on the planning, conduct and reporting of such sensitivity analyses, even in standard settings. In this talk we report on such guidance developed by TG1 to help ensure robust study conclusions, with a check list to aid in reporting these types of sensitivity analyses.

We next list broader issues considered in collaboration between the missing data and causal inference topic groups. In cancer trials, missingness may coincide with intercurrent events and hit a non-positivity boundary. We discuss how recording reasons of missingness as a design feature can support more directed sensitivity analyses. We illustrate our findings and solutions, implementing several strategies in an observational study and a clinical trial with health-related quality of life outcomes. By thus supporting the uptake of sensitivity analyses we seek to contribute to improved research quality, transparency and reproducibility.

Session 2A

The importance of methodology and transparency in predictive AI

Gary Collins, University of Birmingham, UK

The transformative potential of predictive AI in healthcare hinges on building trust, which is predicated on transparency across the entire prediction model lifecycle. This talk will address the critical importance of transparency for fostering trustworthy AI in healthcare, from the initial stages of model development through rigorous evaluation in clinical trials. I will explore how predictive AI studies can obscure biases, hinder clinical adoption, and undermine patient safety. To combat these challenges, I will highlight some best practices for transparent AI development and evaluation, including the implementation of key reporting guidelines such as TRIPOD+AI. Furthermore, I will emphasize how they play a crucial role in reproducibility, critical appraisal, and open science in ensuring the robustness and reliability of AI-driven healthcare solutions. I will discuss the importance of openly sharing code, data, and model parameters, and enabling independent validation. By advocating for these principles, researchers can pave the way for a future where AI in healthcare is not only innovative but also ethically sound and demonstrably trustworthy.

From traditional statistical modelling to machine learning algorithms: a perspective from the STRATOS TG9 high-dimensional data viewpoint

Riccardo De Bin (University of Oslo, Norway); Federico Ambrogi (University of Milan and IRCCS Policlinico San Donato, Italy); Lara Lusa (University of Primorska and University of Ljubljana, Slovenia); Jörg Rahnenführer (Technische Universitat Dortmund, Germany); Willi Sauerbrei (Medical Center - University of Freiburg, Freiburg, Germany)

Abstract: The rise of machine learning, driven by technological and algorithmic advances, has brought with it the promise of data-driven solutions to a wide range of problems, particularly when high-dimensional data from multiple sources are available. For example, in the field of precision medicine, information from various omics data can be integrated to identify specific disease subtypes that may benefit from targeted treatments. The traditional statistical approach to data analysis has been contested by the problem-oriented attitude of machine learning. Nowadays, the classical view of the two cultures, although still fascinating, needs to be replaced by a new vision in which inference and algorithms coexist to address modern challenges: Statistical modelling and machine learning represent a continuum that may help analysts in different

contexts or in specific problems. In this talk, we provide a perspective on this issue from the STRATOS TG9 High-dimensional data viewpoint, focusing on similarities, differences, and common requirements (e.g., reproducibility, reporting) between the two approaches.

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

Ben Van Calster (KU Leuven, Leuven, Belgium**)**, Ewout W Steyerberg (University Medical Center Utrecht, the Netherlands) on behalf of STRATOS TG6.

Selecting appropriate performance measures is essential for clinical AI predictive models intended to support medical decisions. The use of suboptimal performance measures may lead to implementing AI models that suggest detrimental clinical decisions.

We illustrate this challenge using the ADNEX model, which predicts the probability of malignancy in women with an ovarian tumor. We externally validate ADNEX (n=894) and a recalibrated version tailored to the population at hand. Although recalibration improved the model, some measures suggested worse performance (e.g. F1).

We evaluate the merits of classic and contemporary performance measures and plots for models with a binary outcome. We consider five performance domains (discrimination, calibration, overall, classification, utility) and identify two key characteristics for selecting a performance measure: ‘properness’ and ‘focus’. Only 17 of 32 assessed measures possessed both characteristics, while one possessed neither (F1).

Several problematic measures are becoming increasingly popular, making it challenging to improve model validation practices. To provide guidance, we recommend the following measures and plots as essential to report: c-statistic (or AUROC), calibration plot, a clinical utility measure with a decision curve, and a plot showing the distribution of estimated risks.

(Published as: Van Calster B, Collins GS, Vickers AJ, Wynants L, Kerr KF, Barreñada L, Varoquaux G, Singh K, Moons KGM, Hernandez-Boussard T, Timmerman D, McLernon DJ, van Smeden M, Steyerberg EW, for TG6 of the STRATOS initiative. Performance evaluation of predictive AI models to support medical decisions: overview and guidance. Lancet Digital Health 2025;7:100916.)

Designing Prognostic Factor Studies: Key Concepts and Practical Guidance

Peggy Sekula (Medical Center - University of Freiburg, Freiburg, Germany**)**, Suzanne M Cadarette (University of Toronto, Canada), Mitchell H Gail (Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland, USA for topic group 5 of the STRATOS initiative

The investigation of factors that predict clinically relevant outcomes in diseased persons (prognostic factor research) is important to support the development of prognostic tools that can ultimately be used to tailor treatment decisions.

As the output and quality of such studies are in need of improvement, the topic group 'TG5: study design' of the STRATOS initiative has developed a guidance document for general readership, intended to assist the design of studies on prognostic factors. The document also includes a glossary of terms and a list of general aspects to consider when designing a prognostic factor study.

In this presentation, we will provide an overview of the aspects covered in the document and illustrate them using a number of examples.

Peggy Sekula, Inga Steinbrenner, Ulla T Schultheiss, Neus Valveny, Paola Rebora, Susan Halabi, Suzanne M Cadarette, Richard D Riley, Gary S Collins, Willi Sauerbrei, Mitchell H Gail for topic group 5 of the STRATOS initiative. Design aspects for prognostic factor studies. BMJ Open. 2025;15(8): e095065. doi: 10.1136/bmjopen-2024-095065.

Session 2B

Meta-research to improve evidence generation and synthesis

Matthias Briel, University Hospital Basel, Switzerland

Randomized clinical trials (RCTs) are critical for the generation of trustworthy evidence on clinical interventions, and individual participant data meta-analysis (IPDMA) is deemed the 'gold standard' for synthesizing randomized evidence to best inform clinical practice and policy. In this talk, I will present meta-research showing that one out of four RCTs is prematurely discontinued, mostly due to poor recruitment of participants and particularly affecting investigator-initiated RCTs. Results of one third of started investigator-initiated RCTs are not published in peer-reviewed publications or trial registries. Results of prematurely discontinued RCTs even remain unpublished in about 50% of cases. Non-publication of randomized evidence may seriously compromise evidence synthesis

for clinical decision making alongside other issues such as low reporting quality of trials, heterogeneity of trial outcomes and subgroup definitions, or flawed trial data. While IPDMA is able to overcome many of these issues, it has its own challenges. I will present current meta-research on IPDMAs and discuss how the field of IPDMA has changed in the last two decades including the introduction of data sharing policies and the assessment of IPD for integrity issues. Finally, I will outline how prospective collaborative IPDMA can address the need for timely generation of high-quality evidence summaries.

The Problem with Univariable Selection in Regression Modelling — and What to Do Instead

Theresa Ullmann (Medical University of Vienna, Austria), Georg Heinze (Medical University of Vienna, Austria), Franziska Kappenberg (University of Bonn, Bonn, Germany), Marc Henrion (Malawi-Liverpool-Wellcome Trust Clinical Research Programme, Blantyre, Malawi), Daniela Dunkler (Medical University of Vienna, Austria), on behalf of TG2 of the STRATOS initiative

Univariable selection, the screening of variables individually for statistical significance before inclusion in multivariable regression models, is one of the most widely used variable selection strategies in medical research. However, this approach lacks a theoretic justification, since a variable's association with an outcome in isolation does not reflect its role in a multivariable model. As a result, univariable selection can exclude important variables, retain spurious ones, and lead to biased estimates, invalid confidence intervals, and poor predictive performance.

Three key mechanisms explain this failure. Masking occurs when correlated variables obscure each other's true effects, causing important variables to appear non-significant in univariable analyses. Amplification occurs when a variable appears significant in a univariable model only because it is correlated with another variable that has a true effect. Finally, even when variables are uncorrelated, univariable analyses may still fail to detect important variables due to a loss of precision: including multiple relevant variables in a multivariable model reduces the residual error and can reveal predictors that univariable analyses miss.

We illustrate these problems using simulations and two real-world medical datasets (a breast cancer biomarker study and a study of lung health in schoolchildren). Combining univariable screening with subsequent backward elimination does not remedy the problem.

We discuss alternatives to univariable selection and offer concrete recommendations for variable selection tailored to descriptive, predictive, and causal modelling goals.

Guidelines to the design, analysis and interpretation of patient-reported outcomes in cancer clinical trials

Saskia le Cessie (Leiden University Medical Centre, the Netherlands), Doranne Thomassen (Leiden University Medical Centre, the Netherlands), Ahu Alanya (EORTC, Brussels, Belgium), Els Goetghebeur (Ghent University, Belgium)

Since 2019, STRATOS is involved in the international SISAQOL-IMI consortium. This consortium aims to establish international standards in the analysis of patient reported outcomes (PRO) and health-related quality of life data in cancer clinical trials. Last year the SISAQOL-IMI consortium delivered its final recommendations for the design, analysis, interpretation and reporting of PROs in Randomised Clinical Trials (RCT) and Single Arm Studies (SAT). These recommendations can be found on https://www.sisaqol.org/. STRATOS contributed primarily in the work on SAT. In this talk we will give an overview of the results of this work, discussing questions and challenges in the design and analysis of single arm PRO studies.

An overview and categorization of papers published in statistical series of medical journals

Willi Sauerbrei (Medical Center - University of Freiburg, Freiburg, Germany**)** and Gary Collins (University of Birmingham, UK) for the STRATOS initiative

STRATOS compiled a collection of statistical articles published in series of medical journals between 1994 and 2025. These articles are primarily intended to be accessible to clinicians and researchers who have limited familiarity with statistical methodology. Applying a set of inclusion and exclusion criteria, we identified 35 yielding with 1188 articles for consideration. Statistical topics were classified into 36 categories (e.g., bias/systematic errors, confounding, sample size), seven of them with additional subcategories (e.g., Statistical modelling: a: model building, b: sensitivity analyses/check of model assumptions, c: other topics). To ensure consistency, 12 pairs of reviewers double-checked each paper. Members from all STRATOS Topic Groups, along with contributors from some panels assigned each article to up to three categories or recommended its exclusion. In total, 914 articles from 34 journals were categorized.

The primary output of this project is a spreadsheet documenting, for each article, the number of pages, tables, figures, references and the assigned categories. It enables straightforward

searching by topic and provides direct hyperlink access to articles of interest.

This classification resource will support researchers in efficiently identifying methodological literature on a given statistical topic, drawing exclusively from statistical series published in medical journals.