Mini-symposium of the STRATOS initiative at ISCB 2025

Title: Statistical Research needs to improve – on the important roles of simulation studies and guidance for analysis

Organizers: Anne-Laure Boulesteix (Munich, Germany), Willi Sauerbrei (Freiburg, Germany)

Although new biostatistical methods are published at a very high rate, many of these developments are not independently evaluated, raising potential concerns about the accuracy and validity of the results. Similar to the well-known phases of research in drug development, Heinze et al. (2024) propose to identify four phases of methodological research.

In the first session, we will have four talks starting with an introductory presentation of the phases concept, followed by three presentations, each revisiting the development history of a specific important biostatistical method, in light of this concept. These three talks will aim at illustrating what phases of methods' development and evaluation were considered and how they were implemented. Together, they may contribute to a further refinement of the phases of methodological research and stimulate discussions around these pivotal issues.

In the second session we will have four talks from TG3 (Carsten Schmidt), TG5 (Rima Izem), the open science panel (Sabine Hoffmann) and about a joint project of TGs 2 and 4 (Aris Perperoglou).

Program

Mini-Symposium 3, *August 28th, 2025*

Session 1 (09:00-10:30): Phases of methodological research; Chair: Anne-Laure Boulesteix

09:00-09:20 Georg Heinze (Medical University of Vienna, Austria): **How Biostatistical Methods Mature: Understanding the Four Phases of Methodological Research**

09:20-09:40 Michal Abrahamowicz (McGill University, Montreal, Canada): **Phases of development of the Weighted Cumulative Exposure modeling**

09:40-10:00 Willi Sauerbrei (Medical Center - University of Freiburg, Germany): **Phases of development of the Multivariable Fractional Polynomial Interaction (MFPI) approach**

10:00-10:20 Ewout Steyerberg (University Medical Center Utrecht, Netherlands): **Will Net Benefit trump Net Reclassification Index as a measure for incremental value of markers in prediction models? A historical perspective**

10:20-10:30 ***Discussion***

Session 2 (11:00-12:30): Talks from TGs and panels; Chair: Willi Sauerbrei

11:00-11.23 Rima Izem (Novartis, Basel, Switzerland) for TG5: **Mission impossible? Specifying target estimands for long-term risks and benefits of novel therapies**

11:23-11:45 Carsten Oliver Schmidt (University of Greifswald, Germany) for TG3: **An overview on recent works and activities of the STRATOS topic group TG3 "Initial data analysis"**

11:45-12:08 Sabine Hoffmann (Ludwig-Maximilians-Universität Munich, Germany) for the open science panel: **An overview and recent developments of the STRATOS Open Science panel**

12:08-12:30 Aris Perporoglou (GSK, London, UK) for TG2/TG4: **Adjusting for Covariate Measurement Error in Non-Linear Regression: Comprehensive Phase 2 Results from the STRATOS TG2-TG4 Study**

Program abstracts:

Session 1

How Biostatistical Methods Mature: Understanding the Four Phases of Methodological Research

Georg Heinze[1], Anne-Laure Boulesteix[2], Michael Kammer[3], Tim P. Morris[4], Ian R. White[4]

[1]Medical University of Vienna, Center for Medical Data Science, Institute of Clinical Biometrics, Vienna, Austria

[2]LMU Munich, Institute for Medical Information Processing, Biometry and Epidemiology, Munich, Germany

[3]Medical University of Vienna, Department of Medicine III, Division of Nephrology, Vienna, Austria

[4]UCL, MRC Clinical Trials Unit, Institute of Clinical Trials and Methodology, London, UK

Similar to the well-known phases of research in drug development, Heinze et al. [1] identified four phases of methodological research in biostatistics. The initial phase (I) centers on the development of a novel method. It typically raises the rationale for the method's relevance and novelty, and includes theoretical justifications or formal mathematical proofs. Basic illustrations or toy examples may be contained, but comprehensive empirical evaluation is usually not yet considered. This phase is often limited to a single publication presenting the new idea.

In Phase II, an initial evaluation in a controlled and limited simulation setting is performed. Typically, the method's properties are tested under ideal or simplified conditions, and using well-defined, specific data structures. This phase may include an illustrative real data example and provide a first software implementation, but generalizability is not yet the focus. Many journal articles with biostatistical contributions could be assigned to this phase, but few of them make it to the next phase.

Phase III comprises broad evaluations of a method across diverse settings to investigate a method's wider applicability. This usually includes simulation studies or example applications that span over various settings (e.g., different sample sizes, effect sizes, distributions, sometimes even different types of outcome variables). Alternative methods are well-selected based on evidence, and comparisons among methods are often conducted as neutral comparison studies, avoiding or at least disclosing possible biases. This phase helps in identifying strengths and weaknesses of methods across multiple use cases.

The final phase IV provides meta-methodological insights of a method already in use by practitioners (other than its inventors): it further clarifies when and why the method works well or poorly. It may comprise a narrative or systematic review of simulation results, or wide comparative performance studies, sometimes largely extending the originally intended fields of application. After that phase, a method is recognized as mature, enabling recommendations for or against its use in specific contexts or according to potential users' level of statistical knowledge and experience. Finally, guidance documents or tutorials for applied users may emerge.

In this introductory talk I will discuss these and some further aspects of the classification and the impact it may have on different stakeholders, such as methodologist, applied researchers, reviewers and journal editors, and funders and policy makers.

1. Heinze, G., Boulesteix, A. L., Kammer, M., Morris, T. P., White, I. R., & Simulation Panel of the STRATOS initiative (2024). Phases of methodological research in biostatistics-Building the evidence base for new methods. Biometrical journal, 66(1), e2200222. https://doi.org/10.1002/bimj.202200222

Phases of development of the Weighted Cumulative Exposure modeling

Michal Abrahamowicz[1]

[1] McGill University, Montreal, Canada

Weighted Cumulative Exposure (WCE) methodology has been developed to allow for flexible modelling of the cumulative effects of time-varying exposures (TVE) [1]. In time-to-event analyses, the joint impact of past TVE values, for person *i*, is quantified as: $WCE_i(u) = \sum_t w(u-t)[X_i(t)]$, where *u* is the current time when the hazard is assessed, and $X_i(t)$, $t < u$, represent TVE values observed at earlier times. The essential component of the model is the weight function $w(u-t)$ which is estimated using cubic splines and indicates how the importance of the TVE value observed at time *t,* for the hazard at time *u* ($u > t$), varies with time ($u-t$) since it was measured [1]. The WCE modeling, originally developed for Cox proportional hazards analyses [1], has been extended to competing risks, marginal structural models and mixed effects linear modeling of longitudinal changes in a quantitative outcome [2].

The talk will present an overview of the phases of the development and establishing of the WCE methodology, including its consecutive extensions, and validation in simulations. I will discuss how and to what extent our work on the WCE modelling followed the phases identified by Heinze et al in their recent paper on the phases of methodological research in biostatistics [3]. In addition, further phases such as (a) establishing the need for the new methodological development, (b) proof-of-the-concept phase, and (c) refining the estimation and statistical inference, will be outlined.

In this context, three important aspects of real-world applications will be briefly discussed. (i) Firstly, I will emphasize the need to incorporate substantive knowledge, and the related challenges. (ii) Secondly, I will illustrate the ability of the WCE analyses to provide new insights into, and generate new hypotheses about, the underlying biological processes linking the exposure with the outcomes. (iii) I will also outline how some real-world results stimulated new methodological developments, necessary to address additional analytical challenges.

Finally, the need of future research to carry out the additional phase, focusing on neutral simulations to further validate the WCE methodology and compare it with the existing alternative approaches, as recommended in [3], will be briefly presented,

1. Sylvestre M-P & Abrahamowicz M. Flexible modeling of the cumulative effects of time dependent exposures on the hazard. *Statistics in Medicine*. 2009 Nov;28(27):3437-3453.
2. Abrahamowicz M. Assessing cumulative effects of medication use: new insights and new challenges. *Invited Commentary*. *Pharmacoepidemiology and Drug Safety*. 2024 Jan;33(1):e5746. doi: 10.1002/pds.5746.
3. Heinze G., Boulesteix AL, Kammer M., Morris TP, White I. Phases of methodological research in biostatistics. *Biometrical Journal.* 2024.

Phases of development of the Multivariable Fractional Polynomial Interaction (MFPI) procedure

Willi Sauerbrei[1], Patrick Royston[2]

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

[2]MRC Clinical Trials Unit at UCL, Institute of Clinical Trials and Methodology, University College London, London, UK

MFPI is an extension of the well-established Multivariable Fractional Polynomial (MFP) approach to regression modelling. MFPI was formulated in the context of RCTs to look for an interaction with a continuous variable. The linear interaction model is the simplest special case. More generally, the aim is to investigate for an interaction of a categorical variable with a continuous variable in the framework of a regression model [1]. In the clinical context (RCTs), the key component of this method is the continuous treatment effect function (TEF). We used the bootstrap to perform stability analyses of such functions [2]. Our procedure was inspired by the STEPP (Subpopulation Treatment Effect Pattern Plot) approach. The latter was in vogue some 25 years ago to investigate possible interactions in breast cancer research [3]. We compared the approaches in some examples [4].

Regarding selection of the specific functions, we initially suggested four approaches with varying flexibility (FLEX1 to FLEX4). The details are demonstrated in a Stata paper in which we also compared MFPI with STEPP [5]. Using a large simulation study, we showed the advantages of MFPI over categorization-based methods. Regression splines were also considered as competitors and did not yield much better results. No other spline approaches were investigated [6, 7].

We proposed a strategy to average several functions [8] which allowed us to conduct meta-analyses for a continuous variable. Using IPD data from eight RCTs in breast cancer, we illustrated several methodological issues relating to averaging the eight TEF functions [9].

1. Royston, P. and Sauerbrei, W. (2004): A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. Statistics in Medicine, 23:2509-2525.
2. Sauerbrei, W. and Royston, P. (2007). Modelling to extract more information from clinical trials data: On some roles for the bootstrap. *Statistics in Medicine*, *26*(27), 4989-5001.
3. Bonetti, M. and Gelber, R.D. (2000): A graphical method to assess treatment–covariate interactions using the Cox model on subsets of the data, Statistics in Medicine 19: 2595–2609.
4. Sauerbrei, W., Royston, P. and Zapien, K. (2007): Detecting an interaction between treatment and a continuous covariate: a comparison of two approaches. Computational Statistics and Data Analysis, 51: 4054-4063.
5. Royston, P. and Sauerbrei, W. (2009): Two techniques for investigating interactions between treatment and continuous covariates in clinical trials. The Stata Journal, 9: 230-251.
6. Royston, P. and Sauerbrei, W. (2013): Interaction of treatment with a continuous variable: simulation study of significance level for several methods of analysis. Statistics in Medicine, 32(22):3788-3803.
7. Royston, P. and Sauerbrei, W. (2014): Interaction of treatment with a continuous variable: simulation study of power for several methods of analysis. Statistics in Medicine, 33: 4695-4708.
8. Sauerbrei, W. and Royston, P. (2011): A new strategy for meta-analysis of continuous covariates in observational studies. Statistics in Medicine, 30(28):3341-3360.

9. Sauerbrei, W., & Royston, P. (2022). Investigating treatment-effect modification by a continuous covariate in IPD meta-analysis: an approach using fractional polynomials. BMC medical research methodology, 22(1), 98.

Will Net Benefit trump Net Reclassification Index as a measure for incremental value of markers in prediction models? A historical perspective

Ewout Steyerberg[1], Ben Van Calster[2] for STRATOS TG6

[1] University Medical Center Utrecht, Netherlands

[2] KU Leuven, Belgium

Intro: Markers such as lab measurements and omics features hold promise to improve predictions for individual patients. Various measures can be used to quantify the incremental value of such markers. We aim to place two relatively recent measures in historical perspective: Net Benefit (NB) and Net Reclassification Index (NRI).

Methods: The NB was introduced by Vickers & Elkin in 2006 [1], and Net Reclassification Index (NRI) by Pencina et al in 2008 [2]. Both papers have high citations rates (total >4000 and >6000; in 2024: 553 and 295 respectively). Both measures can consider the situation that a reference prediction model is extended with a covariate, either categorical or continuous ('marker extension').

Results: The NB fits in the line of research on utility measures, where true positive (TP) classifications usually are weighted as more important than false positive (FP) classifications. NB is weighted sum of TP and FP, with the weight related to the decision threshold to classify patients as high vs low risk. A related measure is Relative Utility, as proposed by Baker [3].

The NRI is a reclassification measure, where higher risk is an improvement for those with an event, and lower risk for those without an event. For binary classification, the sum of NRI for events and NRI for non-events is equal to the improvement in Youden index (difference in sensitivity plus difference in specificity). Remarkably, Youden index and NB were already described in 1884 in a 1 page paper [4]. The NRI has been criticized for various reasons, including statistically improper behaviour (testing and estimation problems) and fundamental limitations (not accounting for consequences of classifications, which may be context-dependent, and a risk of misinterpretation) [5]. The NRI is equal to NB if the decision threshold is the event rate, so can be considered a simplified case of NB.

Conclusion: NRI and NB arose from different research traditions, which were already defined in 1884. NRI did not go through systematic phases of evaluation and should not be a prime performance measure for the performance of markers to classify patients as low versus high risk. Time will tell whether NB will trump NRI.

1. AJ Vickers, EB Elkin. Decision curve analysis: a novel method for evaluating prediction models. Medical Decision Making 2006: 26 (6), 565-74
2. MJ Pencina, RB D'Agostino Sr, RB D'Agostino Jr, RS Vasan. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 2008: 27(2), 157-72
3. SG Baker. Putting risk prediction in perspective: relative utility curves JNCI 2009:101;1538–42
4. CS Peirce. The numerical measure of the success of predictions. Science, 1884

5. M Leening, M Vedder, J Witteman, M Pencina, M Pencina, E Steyerberg. Net reclassification improvement: Computation, interpretation, and controversies: A literature review and clinician's guide. Ann Intern Med 2014:160,122-31

Session 2

Mission impossible? Specifying target estimands for long-term risks and benefits of novel therapies

Rima Izem[1], Paola Rebora[2], Nicholas Bakewell[3], Mitchell Gail[4], Suzanne Cadarette[5] for TG5

[1] Statistical Methodology, Novartis Pharma AG, Basel, Switzerland

[2] School and Medicine and Surgery, University of Milano-Bicocca, Italy

[3] Health Services Research, University of Toronto, Canada

[4] Biostatistics Branch, National Institutes of Health, Rockville, Maryland, USA

[5] Leslie Dan Faculty of Pharmacy, University of Toronto, Canada

The STRATOS Study Design Topic Group (TG5) aims to offer guidance on planning and designing observational studies. Proper planning, informed by subject-matter expertise, ensures that research objectives are clearly defined, clinically relevant, and that the chosen study design is appropriate and valid. Despite its apparent simplicity, flaws in study design are frequently reported, highlighting the need for robust guidance from this subteam.

One TG5 topic of interest includes the review of main challenges in planning clinical trials or observational studies to answer causal questions about the long-term risks and benefits of treatments for chronic conditions. In chronic care, extended exposure to treatments raises questions about long-term safety or effectiveness, necessitating further studies.

The current practice often involves designing studies to compare the initiation of a new treatment with standard care on long-term outcomes. However, the treatment landscape is dynamic. Patients may experience multiple intercurrent events after initiating a treatment, such as dose escalation, switching treatments, therapy gaps, or concurrent treatments, which can influence outcomes. Ignoring these intercurrent events or censoring follow-up at these events allows estimation but muddies the causal inference. Therefore, estimands often focus on quantifying the effect of treatment initiation at the expense of complex exposure patterns.

A potential alternative that TG5 is exploring is to ask cumulative exposure questions at fixed follow-up periods informed by drug utilization patterns in real-world settings.

An overview on recent works and activities of the STRATOS topic group TG3 "Initial data analysis"

Carsten. O. Schmidt[1]**,** Marianne Huebner[2], Lara Lusa[3]

[1]Institute for Community Medicine, University Medicine of Greifswald, Greifswald, Germany

[2]Department of Statistics and Probability, Michigan State University, East Lansing, MI, USA

[3]Department of Mathematics, Faculty of Mathematics, Natural Sciences and Information Technology, University of Primorska, Koper, Slovenia

The key principle of Initial Data Analysis (IDA) is to provide reliable knowledge about the data underlying the main statistical analyses (MDA). The STRATOS topic group TG3 "Initial data analysis" aims to improve awareness of IDA as an important part of the research process and to provide guidance on conducting IDA in a systematic and reproducible manner in pursue of transparent and reproducible science. IDA focuses on the workflow from metadata setup, data cleaning, data screening, data quality assessments, reporting prior to conducting the MDA. This talk will provide an overview on these steps and introduces an international effort to develop a statistical analysis plan template in cooperation with all STRATOS topic groups for observational studies that incorporates a systematic IDA plan.

An overview and recent developments of the STRATOS Open Science panel

Sabine Hoffmann[1] for the Open Science panel

[1] Institute for Medical Information Processing, Biometry, and Epidemiology, Ludwig-Maximilians-Universität Munich, Germany

The scientific community, publishers and funders are increasingly encouraging open science practices with the idea that "scientific knowledge of all kinds, where appropriate, should be openly accessible, transparent, rigorous, reproducible, replicable, accumulative and inclusive" [1]. The STRATOS Open Science panel was funded to promote open science practices by providing guidance on ways to achieve this idea. This talk with give a general overview of the importance of open science practices in the design and analysis of observational studies in biomedical research and then focus on two ongoing projects concerning guidance on data sharing through synthetic data generation and a project that illustrates how to deal with analytical choices ("researcher degrees of freedom") in the analysis of observational studies.

1. Parsons S, Flavio Azevedo F, Elsherif MM, Guay S, Shahim ON,Govaart GH, Norris E, Aoife O'Mahony A, Parker AJ, Todorovic A, et al. A community-sourced glossary of open scholarship terms. *Nature Human Behaviour,* 6(3):312–318, 2022

Adjusting for Covariate Measurement Error in Non-Linear Regression: Comprehensive Phase 2 Results from the STRATOS TG2-TG4 Study

Aris Perperoglou[1], Mohammed Sedki[2], Anne Thiébaut[3], Michal Abrahamowicz[4], Paul Gustafson[5], Victor Kipnis[6], Laurence Freedman[7] on behalf of the STRATOS TG2 & TG4 collaborative groups

[1] GSK, London, UK

[2] Université Paris-Saclay, France

[3] INSERM National Institute of Health and Medical Research, Villejuif, France

[4] McGill University, Montreal, Canada

[5] Department of Statistics, The University of British Columbia, Vancouver, British Columbia, Canada

[6] Biometry Research Group, Division of Cancer Prevention, National Cancer Institute, Bethesda, Maryland, USA

[7] Biostatistics Unit, Gertner Institute for Epidemiology and Health Policy Research, Sheba Medical Center, Tel Hashomer, Israel

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

This blinded, multi-stage simulation project, a collaboration within the STRATOS initiative (Topic Groups 2 and 4), involved a Data Generation and Evaluation team and three Methods teams. These teams applied Bayesian methods, Imputation/Regression Calibration (MI/RC), and Simulation Extrapolation (SIMEX), combined with flexible modelling techniques (B-splines (BS), P-splines (PS), Fractional Polynomials (FP2), and Natural Splines (NS)). Datasets featured a binary outcome, a continuous covariate with classical error (X*), and a replicate substudy. The true non-linear functional form, covariate distribution, error variance, and error distribution were initially withheld. Phase 1 used 5 pilot datasets; Phase 2 expanded to 155 unique datasets by varying sample sizes, measurement error (ME) variance, error distribution (Normal, shifted-Gamma), and true functional forms. Performance was assessed by log Mean Absolute Error (logMAE).

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