**Mini-symposium of the STRATOS initiative at GMDS 2025, September 9th, 2025**

Organizer: Willi Sauerbrei (Freiburg, Germany)

**Program**

09:00-09:30 Willi Sauerbrei (Medical Center - University of Freiburg, Germany): **Introduction and overview of progress of the STRATOS initiative**

09:30-10:00 Gregor Buch (University Medical Center of the Johannes Gutenberg University Mainz, Mainz, Germany): **A systematic review of statistical model selection techniques used to predict Covid-19 health outcomes**

10:00-10:30 Carsten. O. Schmidt (University Medicine of Greifswald, Greifswald, Germany): **Initial Data Analysis is the basis for responsible statistical analyses: Works of STRATOS Topic Group 3**

## Introduction and overview of progress of the STRATOS initiative

**Willi Sauerbrei**[1], Michal Abrahamowicz[2], Marianne Huebner[3], Ruth Keogh[4] for the STRATOS initiative

[1]*Institute for Medical Biometry and Statistics, Medical Center University of Freiburg, Germany*
[2]*Department of Epidemiology and Biostatistics, McGill University, Montreal, Canada*
[3]*Department of Statistics and Probability, Michigan State University, East Lansing, USA*
[4]*Department of Medical Statistics, London School of Hygiene & Tropical Medicine, London, UK*

The STRATOS initiative (https://www.stratos-initiative.org/en.), launched in 2013 at the annual meeting of the International Society for Clinical Biostatistics (ISCB), aims to provide accessible, evidence-based guidance for key topics in the design and analysis of observational studies. Guidance is intended for applied statisticians and other data analysts with varying levels of statistical expertise and experience. While the primary focus is on health sciences research, the content is also applicable in other empirical sciences.
Currently, the STRATOS initiative comprises nine topic groups (TGs) and ten cross-cutting panels coordinating the activities of the initiative and working on issues common to all TG's. By the end of 2024 STRATOS has published 34 articles. In 2017, STRATOS was invited to publish short articles in the Biometrical Bulletin, the newsletter of the International Biometric Society (IBS). Thirty published articles provide an overview of the work and progress of the initiative. In this talk we will discuss some of the topics and recent progress.

## A systematic review of statistical model selection techniques used to predict Covid-19 health outcomes

**Gregor Buch**[1], Michael Kammer[2,3], Marc Y. R. Henrion[4,5], Georg Heinze[2] on behalf of TG2 of the STRATOS initiative

[1]*Preventive Cardiology and Preventive Medicine, Department of Cardiology, University Medical Center of the Johannes Gutenberg University Mainz, Mainz, Germany*
[2]*Medical University of Vienna, Center for Medical Data Science, Institute of Clinical Biometrics*

[3]*Medical University of Vienna, Department of Medicine III, Division of Nephrology and Dialysis*
[4]*Malawi Liverpool Wellcome Research Programme, Blantyre, Malawi*
[5]*Department of Clinical Sciences, Liverpool School of Tropical Medicine, Liverpool UK*

**Background**: Following the outbreak of the Covid-19 pandemic, the scientific community rapidly developed models to predict health outcomes. The Strengthening Analytical Thinking for Observational Studies (STRATOS) initiative's Topic Group 2 (TG2) 'selection of variables and functional forms in multivariable analysis' has initiated a review of the variable and functional form selection techniques used in these publications. It builds on an existing work by Wynants et al. (2020) but focuses on selection approaches. TG2 members hypothesised that, during the health crisis, researchers relied on methods that they considered trustworthy and robust. Therefore, these models provide a valuable opportunity to examine which methods are currently used to select variables and functional forms.

**Method**: A detailed study protocol was developed, including information on the objectives, eligibility criteria, the procedure for identifying a paper's primary model, and an overview of the data extraction process. On this basis, a structured questionnaire was designed to collect detailed information about the modelling strategy and its suitability confirmed by a pilot study. Both documents were approved by TG2 members and registered at the Open Science Framework (https://osf.io/2afuz/) prior to the review process. Articles that had been systematically identified by Wynants et al. (2020) were re-reviewed by 20 independent statistical reviewers. Two reviewers extracted data from each article and resolved discrepancies by consensus. The main interest was in the modelling steps involved in selecting variables, interactions, and functional forms, but descriptive statistics, reported model features, and implementation details were also extracted.

**Results**: Data from 181 regression-based prediction models covering linear, logistic, and time-to-event models were extracted. Considerable variability in model selection approaches was observed, with researchers often combining multiple statistical methods. Unidimensional approaches were frequently combined with multidimensional techniques without clear rationale. Variable selection during multivariable outcome modelling was commonly performed using p-values, backward elimination, or the Least Absolute Shrinkage and Selection Operator (LASSO). Interaction effects and non-linear relationships were rarely considered. If done, splines or multivariable fractional polynomials were used for the latter. Confidence intervals for model coefficients were given in many papers, but the additional uncertainty introduced by model selection was generally ignored. Information on descriptive statistics was generally adequate, while information on statistical modelling required to replicate the results was regularly insufficient. Analysis code that could clarify these aspects was almost never provided. Notably, the limited existing best practice recommendations for modelling were rarely referenced.

**Discussion**: The review shows a broad reliance on ad-hoc modeling strategies combining relatively simple but commonly used modelling strategies, with sub-optimal properties. This underlines the need for efforts to raise awareness of recommended modelling strategies and the importance of increased training with tutorials and examples. To improve reproducibility, a stronger emphasis on sharing of analysis code could be beneficial. Journals could play a crucial role in these aspects by promoting adherence to reporting guidelines. In addition to a comprehensive summary of model selection techniques used in practice, the talk will cover examples of hard-to-comprehend descriptions of statistical methods that illustrate our experiences.

**Initial Data Analysis is the basis for responsible statistical analyses: Works of STRATOS Topic Group 3**

**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*

This talk introduces the works of STRATOS Topic Group 3 (TG3) on Initial Data Analysis (IDA), which aims to establish reliable understanding of study data in pursuit of responsible statistical analyses. Unfortunately, the importance of IDA is still not fully recognized by many principal investigators, analysts, and funding agencies. It deals with all assessment and curation steps undertaken prior to the main data analysis (MDA). Failing to properly understand data properties, data provenance, and their potential impact on research objectives and analytical choices before conducting the MDA may lead to the use of inappropriate statistical methods and false conclusions. Therefore, TG3 has addressed IDA from a range of perspectives. First, we developed a framework on the building blocks of IDA (1), distinguishing six steps in its workflow: metadata setup, data cleaning, data screening, initial data reporting, refining and updating the analysis plan, and reporting IDA in research papers. A subsequent systematic review showed how current research papers fail to sufficiently report IDA (2), thus indicating the need for improved guidance on how to conduct and report IDA. As a result, best practice examples in the context of regression-type analyses with cross-sectional and longitudinal data have been developed (3, 4). Often IDA takes considerable time and resources, an important challenge for analysts. How to prepare and conduct a potentially wide scope of data assessments, related data quality checks (5), and how to draw appropriate conclusions must be carefully planned. Thus, IDA should also be incorporated into statistical analysis plans (SAPs). TG3's latest efforts focus on integrating IDA into SAPs to enhance transparency and reproducibility of statistical analyses.

1. Huebner M, le Cessie S, Schmidt CO, Vach W. A Contemporary Conceptual Framework for Initial Data Analysis. Observational Studies. 2018;4(1):171-92. doi:10.1353/obs.2018.0014
2. Huebner M, Vach W, le Cessie S, Schmidt CO, Lusa L. Hidden analyses: a review of reporting practice and recommendations for more transparent reporting of initial data analyses. BMC Med. Res. Methodol. 2020;20:1-10.
3. Heinze G, Baillie M, Lusa L, et al. Regression without regrets -initial data analysis is a prerequisite for multivariable regression. BMC Med. Res. Methodol. 2024;24(1):178. doi:10.1186/s12874-024-02294-3
4. Lusa L, Proust-Lima C, Schmidt CO, et al. Initial data analysis for longitudinal studies to build a solid foundation for reproducible analysis. PLoS ONE. 2024;19(5):e0295726. doi:10.1371/journal.pone.0295726
5. Schmidt CO, Struckmann S, Enzenbach C, et al. Facilitating harmonized data quality assessments. A data quality framework for observational health research data collections with software implementations in R. BMC Med. Res. Methodol. 2021;21(1):63. doi:10.1186/s12874-021-01252-7