09:00-12:30 Mini Symposium 2 (Room 2)

STRengthening Analytical Thinking for Observational Studies (STRATOS) initiative – recent progress and foci for the future

Organizer: Willi Sauerbrei and Els Goetghebeur in collaboration with the STRATOS Steering Group

09:00-09:23 Willi Sauerbrei (University of Freiburg, Freiburg, Germany) for STRATOS: **On six foci for the future of STRATOS**

09:23-09:46 Carsten Oliver Schmidt (Universität Greifswald, Germany) for TG3: **Building blocks of Efficient Initial Data Analysis and Data Quality Assessments – Best practice examples**

09:46-10:09 Malka Gorfine (Tel Aviv University, Israel) for TG8: **An Overview and Recent Developments in the Analysis of Multistate Processes**

10:09-10:32. Ben van Calster (Leuven University, Belgium) for TG6: **Assessing performance when developing or validating clinical risk prediction models in the era of machine learning**

10:32-11:00 Coffee Break

11:00 - 11:30 Aris Perperoglou (GSK , UK) for TG2/TG4: **Adjusting for covariate measurement error on functional form estimation: design and early results from a blinded, collaborative STRATOS project**

11:30 - 11:53 Els Goetghebeur (Ghent University, Belgium) for TG7: **Causal inference moving forward – embracing joint (dis)appearances**

11:53 - 12:16 Katherine Lee (Murdoch Children's Research Institute, Melbourne, Australia) for TG1**: Current and future initiatives in missing data**

12:16 - 12:30 General Discussion

STRATOS mini-symposium at ISCB 2024, July 25 2024

STRATOS initiative – recent progress and foci for the future

Coordinators:

Willi Sauerbrei
Institute of Medical Biometry and Statistics, Medical Center - University of Freiburg, Freiburg, Germany
wilhelm.sauerbrei@uniklinik-freiburg.de

Els Goetghebeur
Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Belgium
Els.Goetghebeur@ugent.be

Description:
The STRATOS (STRengthening Analytical Thinking for Observational Studies) initiative is a large collaboration of experts in many different areas of biostatistical research. It was launched at a half-day Mini-Symposium at ISCB 2013. The objective of STRATOS is to provide accessible and accurate guidance in the design and analysis of observational studies (www.stratos-initiative.org).

We will present recent progress from some topic groups and discuss foci for the near future. Soon after the ISCB we will have our 3rd general meeting at the Lorentz Center in Leiden, Netherlands. Some of the talks look forward to cross topic groups research and include material from the preparation phase for this meeting.

Program abstracts:

Session 1

STRATOS: Six foci for the next three years
Sauerbrei W , Carpenter J, Abrahamowicz M, van Geloven N, Gustafson P, Huebner M, Keogh R, Shaw P, Goetghebeur E for the STRATOS initiative

In this talk we discuss foci for the next 3 years of STRATOS research with the hope that further colleagues may support the initiative. More details were recently published (Carpenter et al, 2023, Biometric Bulletin 40(4), 7-9; available on the website)

Future foci

1. Simulation studies

Simulation studies are key tools for validating and comparing statistical methods, and hence critical to the development of evidence-based statistical guidance. STRATOS will maintain a focus on simulation studies and prioritize improving their methodology over the coming years.

2. Open science

The importance of open science is evident, but it is an extremely broad topic, and still in its infancy. For some challenges we will work on accessible guidance for making research more transparent, reproducible and hence credible.

3. Initial Data analysis (IDA)

The 'Initial data analysis' TG3 aims to improve awareness of IDA as a critical component of the research process, and develop guidance on conducting IDA in a systematic, reproducible manner. Some issues will be discussed in the talk by CO Schmidt.

4. Machine learning (ML) enhanced statistical methods

While ML methodologies promise quick automated data driven answers to many questions, it is obvious that both ML and established statistical methodologies have their specific strengths and weaknesses. Each could benefit from the insights offered by the other. How to do that best and when is not obvious. We plan to identify the ML enhanced statistical methods that are most important for different TG's, and systematically assess their properties in realistic settings.

5. Estimands in observational data analysis.

The term 'estimand' essentially refers to what is being estimated and for whom. In the trials context, the ICHE9 addendum (ICH, 2019) formally defines it in terms of five components which make for clear targets and more transparent reporting. The insights and benefits which the estimands framework is bringing to trials research are equally needed in

observational studies, where much of the relevant methodological expertise was originally developed. This topic is discussed in a parallel mini-symposium with a contribution from a STRATOS project.

6. More guidance for researchers with limited statistical knowledge and experience

From the beginning, STRATOS highlighted that many methodological developments are not implemented in practice. Lack of guidance on practical issues is presumed to be an important hurdle. Researchers with only basic statistical knowledge and limited experience in using statistical methodology need much more help.

Building blocks of Efficient Initial Data Analysis and Data Quality Assessments – Best practice examples

Carsten Oliver Schmidt, Lara Lusa, Marianne Huebner for TG3

Rigorous statistical analyses require an adequate understanding of the underlying data. Gaining such an understanding is the main goal of Initial Data Analysis (IDA) (1) and data quality assessments (DQA) (2). IDA and DQA overlap strongly, but differ in that the former being more focused on assessing the fulfillment of prerequisites for the intended substantive analysis, whereas the latter has a more generic focus on data properties. Several works provide guidance on the building blocks for comprehensive and efficient implementation of IDA and DQA. These building blocks range from the setup of metadata to the assessment algorithms used for IDA and DQA. This talks provides best practice examples on the conduct of IDA and DQA in the context of observational health studies, using data from the Study of Health in Pomerania (SHIP) and the Survey of Health, Ageing and Retirement in Europe (SHARE). It will be illustrated, how a comprehensive information management supports automated assessments to increase the scope and quality of IDA and DQA related 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. 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

An Overview and Recent Developments in the Analysis of Multistate Processes

Malka Gorfine for TG8

Multistate models offer a powerful framework for studying disease processes and can be used to formulate intensity-based and more descriptive marginal regression models. They also represent a natural foundation for the construction of joint models for disease processes and dynamic marker processes, as well as joint models incorporating random censoring and intermittent observation times. This article reviews the ways multistate models can be formed and fitted to life history data. Recent works on pseudo-values and accommodation of random effects as a method of incorporating a dependence on the process history and between-process heterogeneity are also discussed. The software available to facilitate such analyses is listed.

Assessing performance when developing or validating clinical risk prediction models in the era of machine learning

Ben van Calster, Ewout Steyerberg for TG6.

An abundance of performance measures for clinical risk prediction models have been proposed in the statistical and machine learning literature. We aim to provide an overview of contemporary performance measures for models with binary outcomes, motivated by the assessment of the value of the previously developed ADNEX model to predict whether an ovarian tumor is malignant in external validation data (n=894, 49% malignant tumors).

We consider five domains of model performance. These include overall measures (e.g. Brier score), measures for discrimination (e.g. AUROC), and measures of calibration (e.g. expected calibration error). When supporting a clinical decision for the patient, a decision threshold on the estimated risk is required to define classification as high versus low risk. The 2x2 table of classification versus outcomes can be described with classification measures (e.g. F1) and clinical utility measures (e.g. net benefit). We discuss 32 common performance measures (9 overall, 3 discrimination, 6 calibration, 11 classification, 3 utility). For each performance domain, matching graphical assessments are available.

We define three key desirable characteristics for performance measures: properness (i.e. whether the value of the measure is optimal when the correct risks are used); having an understandable interpretation; and having a clear focus by targeting only one of the five domains. The majority of measures fail for at least one characteristic, while the F1 score fails at all three. All considered classification measures at a given threshold *t* are improper.

A natural requirement is that a performance measure should match the intended use of the model. We discern three common situations. First, when externally validating models that aim to support clinical decision making, it makes sense to assess performance in the following order: discrimination (AUROC), calibration (calibration plot) and clinical utility (net benefit). Second, if a model is merely used for informing/counseling patients about their risk, external validation should focus on calibration. Third, when methodologically comparing multiple models, overall measures are useful. Other measures may be added, if they meet the three key characteristics.

In conclusion, we recommend to consider a limited set of key measures to assess performance aspects in relation to the intended use of a prediction model, focusing on (semi-)proper measures with a clear interpretation and focus.

After the Coffee Break: Session 2

Adjusting for covariate measurement error on functional form estimation: design and early results from a blinded, collaborative STRATOS project

Aris Perperoglou, Paul Gustafson, Michal Abrahamowicz, Victor Kipnis, Mohammed Sedki, Anne Thiébaut, Lawrence Freedman for TG2 - TG4

Introduction

The evaluation and estimation of relationships between outcome variables and covariates measured with error remains a challenge in observational studies. Challenges may be amplified when the true functional form between the covariate and the outcome is suspected to be non-linear. This work outlines the collaboration between two topic groups of the STRATOS initiative: TG2, specializing in the selection of variables and functional forms in multivariable analysis, and TG4, which focuses on

addressing issues related to measurement error and misclassification. The project investigates the performance of methods for estimating complex functional relationships in observational data, where covariates are prone to measurement inaccuracies, specifically targeting the accurate estimation of non-linear relationships between outcome variables and covariates.

Project Design

The project adopts a blinded, multi-stage design to rigorously compare methodological approaches. A Data Generation and Evaluation team produces datasets simulating various functional relationships and measurement error scenarios, but withholds the true underlying model from the Methods teams. Three distinct Methods teams implement different analytic approaches, based on Bayesian methods, Imputation/Regression Calibration methods and SIMEX methods, respectively. For each method, estimation of the functional form based on (i) B-splines, (ii) P-splines and (iii) Fractional Polynomials are investigated, with pre-specified hyperparameters for each approach.

The initial phase of the project involved generating 5 datasets, each comprising realizations of a binary outcome Y and a continuous covariate measured with independent error, alongside pairs of repeat covariate observations for validation purposes in a random subset. The relationship between Y and X was specified as a logistic regression, but the error structure and functional relationship between logit(P(Y=1) and X remained undisclosed to prevent bias in analysis and thus enhance the integrity of the study. Each methods team created their code and returned estimated functional relationships without knowledge of the generating model or the results from any other team's methods.

Early Results and Implications

The Evaluation Team assed the performance of the methods through a comparison of predicted values against undisclosed true values, using mean squared error and other relevant metrics to gauge performance. Initial findings reveal performance disparities across methods (blinded for the study's integrity). In Phase 2 of the project a further 75 datasets have been simulated, varying sample size, functional form and size of measurement error in a systematic manner. These datasets are now being analyzed. The results will offer insights into how factors like sample size, validation study size and spline and covariate functions influence method accuracy This analysis will be important as it will guide the selection and refinement of statistical methods for the final phase of the project that will include several hundred more simulated datasets.

This blinded, collaborative structure fosters an unbiased and efficient evaluation of statistical techniques. Results will contribute to the STRATOS Initiative's broader goal of providing guidance for analyzing observational data. Notably, this project design showcases a model for collaborative, transparent, and rigorous statistical research to address challenges in real-world settings.

TG7: Moving forward – embracing joint (dis)appearances

Els Goetghebeur and Saskia le Cessie for TG7 https://stratos-initiative.org/en/group_7

Since its start, TG7 has presented estimands as a needed focus for any causal effect estimation. `What are we actually estimating' is surprisingly often absent from applied publications [2023, *Lancet Oncol.*, DOI10.1016/S1470-2045(23)00110-9]. It is not straightforwardly derived, however, from how we estimate our target but also depends on what plausible causal assumptions we make a priori. From the various principled answers developed in a setting with sequential point exposures and subsequent continuous outcome, we are now moving to guidance on more complex outcomes. These include right censored survival times and repeated outcome measures while patients are alive. Other intercurrent events may then appear. Already in randomized trials there is much controversy about what constitutes a meaningful causal effect in that case. These issues and much more play in observational studies, for which many causal methods were first developed.

We elaborate in this talk on ongoing and planned work, which looks to collaborate with other topic groups next. We are working on the following:

1. To clarify various causal estimands and estimators we have introduced counterfactual cross world simulations for continuous outcomes as a learning tool. We are finalizing a similar plasmode like effort, starting from an observed case study, for right censored survival outcomes.
2. In the context of the European IMI-SISAQOL project (sisaqol-imi.org) we collaborate with a large consortium to develop guidance for causal effect analysis in (late stage) oncology trials. There, quality of life while alive, as well as survival must be jointly evaluated, often in single arm studies. We define relevant and feasible estimands in that setting and develop corresponding estimators.
3. As a rule, we value causal effects in terms of this joint outcome. While intercurrent events can sometimes be handled by defining composite or (not too) hypothetical outcomes, evaluating treatment policy and outcomes until death is often preferred. When implementing analyses for the joint outcome, many analysis choices must be made. We explain how issues inherent to (single arm) trials as well as cohorts can be more rigorously approached by causal inference methods to allow for target effect estimation under transparent assumptions.

In this talk we describe pitfalls and progress made. We refer to a forthcoming collaboration with TG1 to further analysis in the presence of common missing data patterns in our setting. We look forward to discussions in Thessaloniki and further cross topics work.

TG1: Current and future initiatives in missing data

Katherine J Lee, Els Goetghebeur, James Carpenter on behalf of TG1

The aim of TG1 is to describe the principles for the analysis of partially observed observational data, illustrate potential methods for handling missing data and their application, and provide general guidance on how best to handle missing data across a range of settings. We have previously developed a framework for the handling and reporting of missing data. We are currently expanding this framework to the context of when data are missing dependent on unobserved data. This is exemplified through a worked case study. Future initiatives of TG1 include:

1. conducting a review of journal guidelines for handling missing data in top ranked medical journals with the aim of highlighting key misunderstandings, outlining the key components which we believe are useful to include in author guidance for missing data, and suggesting a template for author guidelines,
2. providing an overview of methods for handling missing data including a discussion of plausibility of needed assumptions, pros and cons of the various approaches and example code for conducting each using a single case study, and
3. evaluating methods to handle missing data in the context of informative drop out and non-positivity, where (nearly) all further data are missing for some categories of participants. We consider a case study on missing quality of life data in a cancer trial with substantial treatment discontinuation and drop-out due to disease progression.

In this talk we describe these initiatives of TG1, in particular also how we are developing collaborations with TG7 to address questions regarding the handling of missing data in causal inference.

General discussion