MEDICAL UNIVERSITY
OF VIENNA

Wiener Biometrische Sektion
der Internationalen Biometrischen Gesellschaft
Region Österreich – Schweiz

Einladung zum Herbstseminar

SAMBA – Statistical Model Building Strategies for Cardiologic Applications – New results and future challenges

We cordially invite you to the workshop **Statistical Model Building Strategies for Cardiologic Applications (SAMBA) - New results and future challenges** on November 23th and 24th in Vienna. The workshop is organized by the PIs of the international joint project SAMBA in cooperation with the Wiener Biometrische Sektion (WBS) and the Center for Medical Data Science (CeDAS) at the Medical University of Vienna. We are delighted to welcome Prof. Dr. **Tobias Pischon** from Berlin, Germany, and Prof. Dr. **Maarten van Smeden** from Utrecht, Netherlands, as keynote speakers. We acknowledge the support by the Austrian Science Funds (FWF) and the Deutsche Forschungsgemeinschaft (DFG) through the joint project SAMBA, grants I4739-B and RA-2347/8-1, respectively.

The workshop will cover two half days. On the afternoon of November 23th our focus will be on the **intersection of epidemiology and biostatistics**. Prof. Pischon will talk on the *Interaction of epidemiology with biostatistics within the large prospective cohort study NAKO with emphasis on cardiovascular diseases* followed by a panel discussion on *Complex longitudinal studies: Needs and challenges in the interaction epidemiology – biostatistics*. The morning of November 24th is dedicated to **The power and risks of algorithmic model building**, where Prof. van Smeden will give his talk on *Rage against the machine learning*. The participants of the SAMBA project will also present issues and opportunities of statistical model building strategies with a focus on cardiological applications.

The participation in the workshop is free of charge and open to any national and international participants of any scientific discipline interested in the interaction of biostatistics, epidemiology, machine learning and cardiology.

We are looking forward to welcoming you in Vienna in November!

Daniela Dunkler, Georg Heinze, Heiko Becher, Geraldine Rauch (organizing committee)

Please register by 9 November 2023 at https://biometrician.github.io/SAMBA/. You can also find more information about the workshop there.

Thursday, 23.11.2023

Topic of the day **Epidemiology and biostatistics: we need each other**

Venue: Medical University Vienna, Hörsaalzentrum (building 4 on the attached plan), Hörsaal 3

Start	End	Topic	Presenter
13:00	13:05	Introduction	Daniela Dunkler & Heiko Becher
13:05	13:25	**Causal model building in the context of cardiac rehabilitation: A systematic review and a case study**	Daniela Dunkler
13:25	14:15	Keynote talk: **The German National Cohort (NAKO): Opportunity, challenges, and interactions with biostatistics** (see Section 4.0.1)	Tobias Pischon
14:15	14:45	Panel discussion: **Complex longitudinal studies: Needs and challenges in the interaction between biostatistics and epidemiology**	*Panelists:* Tobias Pischon, Heiko Becher, Ulrike Grittner, Dörte Huscher, Maarten van Smeden, Daniela Dunkler, Georg Heinze
14:45	15:15	COFFEE BREAK	
15:15	15:35	**Phase IV study of joint models for longitudinal and time to event data**	Jil Kollmus-Heege
15:35	15:55	**Using background knowledge from previous studies in model building: the good, the bad and the ugly**	Lorena Hafermann
15:55	16:15	**Correlated predictors: is variable omission a good default solution?**	Mariella Gregorich
16:15	16:20	Closing	Daniela Dunkler

Friday, 24.11.2023

Topic of the day **The power and risks of algorithmic model building**

Venue: Medical University Vienna, Spitalgasse 23, Jugendstilhörsaal

Start	End	Topic	Presenter
9:00	9:10	Introduction	Georg Heinze
9:10	9:30	**Model building using statistical and machine learning approaches: Comparison of performance and explainability**	Christine Schilhart-Wallisch
9:30	10:30	Keynote talk: **Rage against the machine learning**	Maarten van Smeden
10:30	11:00	COFFEE BREAK	
11:00	11:20	**Variable selection methods for linear and logistic regression: A comparison of approaches**	Theresa Ullmann
11:20	11:40	**Reliable confidence intervals after variable selection: does it work in practice?**	Michael Kammer
11:40	12:00	**Variable selection: Improving on the bad habit to ignore selection decisions in model inference**	Nilufar Akbari
12:00	12:05	Closing	Georg Heinze

Abstracts of keynote talks

The German National Cohort (NAKO): Opportunity, challenges, and interactions with biostatistics

Tobias Pischon

The German National Cohort (NAKO) is a large, multidisciplinary, population-based cohort study that aims to investigate the development and etiology of major chronic diseases, identify risk factors and enhance early detection and prevention of diseases. Between 2014 and 2019, a total of 205,217 persons aged 20-74 years was recruited and examined at 18 study centers across Germany. Whole-body 3T magnetic resonance imaging (MRI) readings were performed on 30,861 participants at 5 study centers. Participants are followed-up for incident diseases via questionnaires, health records and disease registries. In addition, a mortality follow-up has been established. In addition, all study participants are being invited for a first re-examination, similar to the baseline program, between 2019 and 2024, and a second re-examination between 2024 and 2028. Interested scientists can use the web-based NAKO TransferHub (transfer.nako.de) to obtain insight into data dictionary and study population, and to apply for data or biological samples, which are made available on the basis of a Use and Access Policy. In my presentation I will provide an overview of the German National Cohort and give examples of opportunity and challenges with a focus on interactions with biostatistics.

Rage against the machine learning

Maarten van Smeden

Medicine appears to be at the start of a new era due to the many promising developments in AI and machine learning (ML). Part of that promise is that ML will improve upon traditional statistical modelling. Another promise is that ML will outperform medical doctors. In this talk I will highlight a few medical settings where ML seems to have lived up to its promises, and some where it has not. A couple of challenges for fair comparisons between machine learning, traditional statistical modelling and doctors will be identified and discussed. Finally, I will speculate on the future role of machine learning, traditional statistical modelling and medical doctors.

Abstracts of other talks

Causal model building in the context of cardiac rehabilitation: A systematic review and a case study

Nilufar Akbari, Georg Heinze, Geraldine Rauch, Ben Sander, Heiko Becher & *Daniela Dunkler*

Randomization is an effective design option to prevent bias from confounding in the evaluation of the causal effect of interventions on outcomes. However, sometimes randomization is not possible, making subsequent adjustment for confounders essential to obtain valid results. Several methods exist to adjust for confounding, with multivariable modeling being among the most widely used. The main challenge is to determine which variables should be included in the causal model and to specify appropriate functional relations for continuous variables in the model. While the statistical literature gives a variety of recommendations on how to build multivariable regression models in practice, this guidance is often unknown to applied researchers. We set out to investigate the current practice of explanatory regression modeling to control confounding in the field of cardiac rehabilitation, for which mainly non-randomized observational studies are available. In particular, we conducted a systematic methods review to identify and compare statistical methodology with respect to statistical model building in the context of the existing recent systematic review CROS-II, which evaluated the prognostic effect of cardiac rehabilitation.

Phase IV study of joint models for longitudinal and time to event data

Jil Kollmus-Heege

In recent years, there has been an increasing use of joint models to describe the relationship between a longitudinally measured biomarker and the time to an event. The current literature suggests that joint models are preferable to conventional methods as they were shown to be unbiased in such settings. However, it is important to conduct further investigation into this relatively new and complex approach to fully understand its pitfalls and to identify what to watch out for when using it. Therefore, we compare joint models to the time-varying Cox model and the two-stage approach in multiple simulation studies. These include basic simulations as well as more complex ones based on actual observational data.

Model building using statistical and machine learning approaches: Comparison of performance and explainability

Christine Schilhart-Wallisch, Asan Agibetov, Daniela Dunkler, Maria Haller, Matthias Samwald, Georg Dorffner & Georg Heinze

While boundaries between classical statistical model building (SMB) and machine learning (ML) are fluid, both modeling cultures have been used to develop cardiovascular prediction models based on typical predictors such as total cholesterol, blood pressure, sex or age. We reanalyzed data from a large registry of 1.5 million participants in a national health screening program. We developed analytical strategies based on SMB (including nonlinear functional forms and interactions) or on ML to predict cardiovascular events within 1 year from health screening. This was done for the full data set and with gradually reduced sample sizes.

Predictive performance was similar if the models were derived on the full data set, whereas smaller sample sizes favored simpler models. Visualized predictor-risk relations differed between SMB and ML, even with large sample sizes. Hence, the supposed role of cardiovascular predictors in prognosis can depend on how a risk prediction model was developed.

Using background knowledge from previous studies in model building: the good, the bad and the ugly

by *Lorena Hafermann*, Heiko Becher, Carolin Herrmann, Nadja Klein, Michael Kammer, Georg Heinze & Geraldine Rauch

The question of which variables to include in a statistical model is of key importance. Thereby, the integration of background knowledge into the modelling process is an issue that is gaining increasing interest. Often background knowledge is based on results from previous studies. Its quality can vary substantially, particular with regard to the methods used to select variables, raising concerns about its utility for model building in a new study. We found that in the context of descriptive and explanatory linear regression models, background knowledge from previous studies that employed inappropriate model building strategies such as univariable selection distorted the specification of an appropriate predictor set. We also investigated if prediction models estimated by Random Forests benefitted from the incorporation of appropriate background knowledge on the set of predictors. Surprisingly, there was no benefit in terms of discrimination, but the calibration of the model improved by well-informed restriction of the candidate predictor set. Therefore, we recommend to use background knowledge with care and to critically appraise the methodological quality of any previous studies from which it was derived.

Correlated predictors: is variable omission a good default solution?

by *Mariella Gregorich*, Susanne Strohmaier, Daniela Dunkler & Georg Heinze

The development of statistical models is often conducted without adequate consideration for the research objective, leading to model misspecification and wrong interpretation. In this study, we investigate commonly employed practices in applied regression analysis when dealing with highly correlated predictors whilst challenging the conventional approach of eliminating one or more correlated variables from the regression model. Instead, we emphasize that the adequate handling of correlated predictors should be tailored to the conceptual research objective: descriptive, explanatory or predictive modelling. Through a range of empirical data examples, we illustrate the limitations of diagnostic statistical tools in guiding data analysts in model building and propose alternative strategies to handle high correlation that align with the research objectives.

Reliable confidence intervals after variable selection: does it work in practice?

by *Michael Kammer*, Daniela Dunkler, Stefan Michiels & Georg Heinze

Biomedical data analysis relies heavily on variable selection for regression models. However, classical statistical frequentist theory, which assumes a fixed set of model covariates, does not cover inference post-selection. We compared selective inference methods for Lasso and adaptive Lasso submodel parameters: sample splitting, selective inference conditional on Lasso selection (SI), and universally valid post-selection inference (PoSI). Using R software packages, we evaluated their selective confidence intervals through simulations mirroring typical biomedical data. We also applied these methods to a real dataset. SI method showed generally acceptable frequentist properties but failed to achieve claimed selective coverage levels, especially with adaptive Lasso. Sample splitting is an alternative if simplicity is favored over efficiency. PoSI was extremely conservative and is suitable only when few predictors undergo selection or avoiding false positive claims is crucial. In summary, selective inference aids in assessing uncertainties in the importance of selected predictors for future applications.

Variable selection methods for linear and logistic regression: A comparison of approaches

by *Theresa Ullmann*, Christine Schilhart-Wallisch, Lorena Hafermann, Georg Heinze & Daniela Dunkler

Data-driven variable selection is frequently performed in statistical modeling, i.e., when modeling the associations between an outcome and multiple independent variables. Variable selection may improve the interpretability, parsimony and/or predictive accuracy of a model. Yet variable selection can also have negative consequences, such as false exclusion of important variables or inclusion of noise variables, biased estimation of regression coefficients, underestimated standard errors and invalid confidence intervals, as well as model instability. Few large-scale simulation studies have neutrally compared data-driven variable selection methods with respect to their consequences for the resulting models. We present results from a simulation study that compares different variable selection methods: forward selection, (augmented) backward elimination, univariable selection, and penalized likelihood approaches (Lasso, relaxed Lasso, adaptive Lasso). In the interpretation of the results, we put a specific focus on which estimands and performance criteria are relevant for which modeling goals (descriptive, explanatory or predictive modeling).

Variable selection: Improving on the bad habit to ignore selection decisions in model inference

by *Nilufar Akbari* & Georg Heinze

For regression models, methods for estimating the variance for the regression coefficients a set of pre-specified predictors are well-established. However, when the set of predictors is based on data-driven variable selection, there is no widely accepted method to account for selection uncertainty. Since it must be assumed that not selecting a variable into a model is associated with uncertainty, variance estimates of the coefficients of the unselected variables are desirable. The idea of this study is to use a modified sandwich variance estimator for this task. Hereby, the robust variance (sandwich) formula is applied to all candidate variables, including variables that were not selected into the final model. We conducted a simulation study to investigate the method and compare its performance to two model-based variance methods and two bootstrap based methods.

Affiliations of presenters, panelists and chairs

Tobias Pischon	Charité Universitätsmedizin Berlin & Max-Delbrück-Centrum für Molekulare Medizin (MDC) Molecular Epidemiology Research Group Berlin, Germany
Maarten van Smeden	University Medical Center Utrecht Julius Center, dept. Epidemiology & Health Economics Utrecht, The Netherlands
Heiko Becher	Universitätsklinikum Heidelberg & Heidelberger Institut für Global Health Unit of Epidemiology and Biostatistics Heidelberg, Germany
Nilufar Akbari Ulrike Grittner Lorena Hafermann Dörte Huscher Jil Kollmus-Heege	Charité Universitätsmedizin Berlin Institut für Biometrie und Klinische Epidemiologie (IBIKE) Berlin, Germany
Christine Schilhart-Wallisch	Agentur für Gesundheit und Ernährungssicherheit Vienna, Austria
Michael Kammer	Medical University of Vienna Center for Medical Data Science, Institute of Clinical Biometrics & Department for Medicine III, Division of Nephrology Vienna, Austria
Daniela Dunkler Mariella Gregorich Georg Heinze Theresa Ullmann	Medical University of Vienna Center for Medical Data Science Institute of Clinical Biometrics Vienna, Austria