Talks in the seminar series of the SFB 876

Thursday, March 22nd, 14:15-17:15 in lecture hall OH14 E23

"High-dimensional data: Design and Analysis"

March 20-23, 2018, TU Dortmund University

1st Workshop of **Topic group 9 High-dimensional data (TG9) of the STRATOS initiative** together with the **Collaborative Research Center (CRC) SFB 876** – Providing Information by Resource-Constrained Data Analysis at TU Dortmund University

14:15-14:30: Willi Sauerbrei (University of Freiburg, Germany):

Short introduction of the STRengthening Analytical Thinking for Observational Studies (STRATOS) initiative

The validity and practical utility of observational medical research depends critically on good study design, excellent data quality, appropriate statistical methods and accurate interpretation of results. Statistical methodology has seen substantial development in recent times, unfortunately often ignored in practice. Part of the underlying problem may be that even experts (whoever they are) do often not agree on potential advantages and disadvantages of competing approaches. Furthermore, many analyses are conducted by applied researchers with limited experience in statistical methodology and software. The lack of guidance on vital practical issues discourages them from using more appropriate methods. Consequently, analyses reported can be flawed, casting doubt on their results and conclusions.

The main aim of the international STRATOS initiative is to develop guidance for researchers with different levels of statistical knowledge. Currently there are nine topic groups on study design, initial data analysis, missing data, measurement error, variable and function selection, evaluating test and prediction models, causal inference, survival analysis, and high dimensional data. In addition, the initiative has ten crossing cutting panels. We will give a short introduction of the initiative. More information is available on the website (http://stratos-initiative.org) and in the first paper (Sauerbrei et al (2014), Statist Med 33: 5413-5432).

14:30-15:15: Lisa McShane (NIH, USA):

Analysis of high-dimensional Data: Opportunities and challenges

"Big data," which refers to data sets that are large or complex, are being generated at an astounding pace in the biological and medical sciences. Examples include electronic health records and data generated by technologies such as omics assays which permit comprehensive molecular characterization of biological samples (e.g., genomics, transcriptomics, proteomics, epigenomics, and metabolomics), digital and molecular imaging technologies, and wearable devices with capability to collect real-time health status and health-related behaviors. Big data may be characterized by "large n" (number of independent observations or records) and/or "large p" (number of dimensions of a measurement or number of variables associated with each independent record). Either large n or p may present difficulties for data storage or computations, but large p presents several particularly interesting statistical challenges and opportunities and is the focus of High-dimensional Data Topic Group (TG9) within the STRATOS initiative.

Many types of high-dimensional data in the biomedical field (e.g., generated from omics assays or by imaging) require pre-processing prior to higher level analyses to correct for artifacts due to technical biases and batch effects. Statistical pre-processing methods may require modification, or novel approaches may be needed, as new technologies emerge. Visualization and exploration of data in high dimensions is also challenging, necessitating development of novel graphical methods, including approaches to integrate high-dimensional data of different types such as DNA mutation calls and expression levels of genes and protein. Additionally, data dimension reduction, for which many methods exist, may be needed for ease of interpretation or as an initial step before proceeding with downstream analyses such as prediction modeling.

Many discovery studies have as their goal identification of biological differences between groups of biological specimens, patients, or other research subjects. When those differences may occur in any of thousands of measured variables, standard approaches that control family-wise error (e.g., Bonferroni adjustment) are generally too stringent to be useful. The explosion of high-dimensional data has encouraged further development of approaches that control expected or actual false discovery number or proportions. Analysts need to appreciate what criteria these methods control and what assumptions are required.

Traditional classification and prediction methods may become computationally infeasible or unstable when the number of potential predictor variables is very large. Penalized regression methods and a variety of machine learning methods have been introduced into routine statistical practice to address these challenges. However, great care is needed to avoid overfitting models in high-dimensions. Methods such as cross-validation or other resampling methods can be used to provide realistic assessments of model performance and detect overfitting; frequent occurrence of overfit models based on high-dimensional data in the published literature suggests that more education is needed on proper model performance assessment.

More research to develop new approaches for analysis of high-dimensional data is clearly needed. Before adoption, performance of new methods should be adequately assessed on real and simulated data sets. Some methods previously developed for use on data of substantially lower dimension might also require reassessment to ensure that their acceptable performance is maintained in high dimensions. How to simulate realistic data in high dimensions is a research topic in itself.

Growth of big data is already outpacing the increase in the number of individuals knowledgeable in how to manage and analyze these data. The goal of the High-dimensional Data Topic Group (#9) of STRATOS is to educate researchers, including statisticians, computational scientists and other subject matter experts, on proper design and analysis of studies reliant on high-dimensional data, and also to stimulate development of new and improved methods for application to big data. Success in meeting the demand for big data analytic methods will require unprecedented levels of collaboration among all parties engaging in big data-based research.

15:15-15:45 Tomasz Burzykowski (Hasselt University, Belgium)

A bird’s eye view on processing and statistical analysis of 'omics' data

Technologies used to collected experimental “omics” data share several important features: they use sophisticated instruments that involve complex physical and biochemical process; they are highly sensitive and can exhibit systematic effects due to time, place, reagents, personnel, etc.; they yield large amounts (up to millions) of measurements per single biological sample; they produce highly structured and complex data (in terms of correlation, variability, etc.). The features pose various practical challenges. For instance, sensitivity to systematic effects can compromise reproducibility of the findings if experiments are repeated in different laboratories. There are also challenges for the statistical analysis of the data. In the presentation we will provide an overview of and illustrate the common points that one may need to keep on mind when attempting to analyze an “omics” dataset.

15:45-16:00 Discussion and break

16:00-16:30 Riccardo de Bin (University of Oslo, Norway):

Strategies to derive combined prediction models using both clinical predictors and high-throughput molecular data

In biomedical literature, numerous prediction models for clinical outcomes have been developed based either on clinical data or, more recently, on high-throughput molecular data (omics data). Prediction models based on both types of data, however, are less common, although some recent studies suggest that a suitable combination of clinical and molecular information may lead to models with better predictive abilities. This is probably due to the fact that it is not straightforward to combine data with different characteristics and dimensions (poorly characterized high-dimensional omics data, well-investigated low-dimensional clinical data). Here we show some possible ways to combine clinical and omics data into a prediction model of time-to-event outcome. Different strategies and statistical methods are exploited.

16:30-17:00 Willi Sauerbrei (University of Freiburg, Germany):

Guidance for the selection of variables and functional form for continuous variables – Why and for whom?

During recent times, research questions have become more complex resulting in a tendency towards the development of new and even more complex statistical methods. Tremendous progress in methodology for clinical and epidemiological studies has been made, but has it reached researchers who analyze observational studies? Do experts (whoever they are) agree how to analyze a study and do they agree on potential advantages and disadvantages of competing approaches?

Multivariable regression models are widely used in all areas of science in which empirical data are analyzed. A key issue is the selection of important variables and the determination of the functional form for continuous variables. More than twenty variable selection strategies (each with several variations) are proposed and at least four approaches (assuming linearity, step functions (based on categorization), various types of spline based approaches and fractional polynomials) are popular to determine a functional form. In practice, many analysts are required de facto to make important modelling decisions. Are decisions based on good reasons? Why was a specific strategy chosen? What would constitute a ‘state-of-the-art’ analysis?

Considering such questions we will argue that guidance is needed for analysts with different levels of statistical knowledge, teachers and many other stakeholders in the research process. Guidance needs to be based on well designed and conducted studies comparing competing approaches. With the aim to provide accessible and accurate guidance for relevant topics in the design and analysis of observational studies the international STRengthening Analytical Thinking for Observational Studies (STRATOS) Initiative (http://stratos-initiative.org) was recently founded. More about issues mentioned is given in the short summary of topic group 2 ‘Selection of variables and functional forms in multivariable analysis’ in a paper introducing the initiative and its main aims (Sauerbrei et al (2014), Statist Med 33: 5413-5432).

17:00-17:15 Discussion