STRengthening Analytical Thinking for Observational Studies (STRATOS):

Introducing the Topic Group on Selection of Variables and Functional Forms in Multivariable Analysis (TG2)

Aris Perperoglou[1], Georg Heinze[2], Willi Sauerbrei[3] on behalf of STRATOS TG2

[1] Department of Mathematical Sciences, University of Essex, UK

[2] Section for Clinical Biometrics, Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Austria

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

The Biometric Bulletin has recently introduced its readership to the STRATOS initiative and described the activities of the Topic Groups on Missing Data (TG1), Measurement Error (TG4) and on Initial Data Analysis (TG3). This series now continues with an introduction to TG2, dealing with selection of variables and functional forms in multivariable analysis.

The members of the Topic Group are Georg Heinze, Aris Perperoglou and Willi Sauerbrei (Joint Chairpersons), Michal Abrahamowicz, Heiko Becher, Harald Binder, Daniela Dunkler, Frank Harrell, Geraldine Rauch, Patrick Royston and Matthias Schmid.

In multivariable analysis, it is common to have a mix of binary, categorical (ordinal or nominal) and continuous variables, which may be related to an outcome. While TG6: Evaluating Diagnostic Tests and Prediction Models focuses on the goal of predicting the outcome as accurately as possible, the main focus of TG2 is to identify explanatory variables and gain insight into their individual and joint relationship with the outcome. Two of the (interrelated) main challenges are: *(i)* selection of variables for inclusion in a multivariable explanatory model, and *(ii)* choice of the functional forms for continuous variables (Harrell 2015, Sauerbrei et al. 2007)

(i) In practice, multivariable models are often built through a combination of

- a *priori* inclusion of well-established explanatory variables of the outcome of interest, and
- a *posteriori* selection of additional variables, based on data-dependent procedures and criteria such as statistical significance or goodness-of-fit measures.

Although there is a consensus that all of the many suggested model building strategies have weaknesses (Miller 2002), opinions on the relative advantages and disadvantages of particular strategies differ considerably.

(ii) The effects of continuous predictors are typically modeled by either categorization, which raises issues as the number of categories, cut-point values, implausibility of the resulting step-function relationships, local biases, power loss, or invalidity of inference in case of data-dependent cut-points (Royston and Sauerbrei 2008), or by using their original form assuming linear relationships with the outcome, or by performing a simple transformation (e.g. logarithmic or quadratic). Often, however, the reasons for choosing such conventional representation of continuous variables are not discussed, and the validity of the underlying assumptions is not assessed.

To address these limitations, statisticians have developed flexible modeling techniques based on various types of smoothers, including fractional polynomials (Royston and Altman 1994, Royston and Sauerbrei 2008) and several 'flavours' of splines, e.g., restricted regression splines (Harrell 2015), penalized regression splines (Wood 2006), smoothing splines (Hastie and Tibshirani 1990) and p-splines (Eilers and Marx 1996). For multivariable analysis, these smoothers have been incorporated in generalized additive models.

Many issues still exist for each of the following parts: (i) selection of variables, *(ii)* selection of functional forms for continuous variables and *(iii)* their combination. Practical guidance is urgently needed, necessitating extended investigations of analytical properties and systematic comparisons between alternative methods. TG2 has started several projects: (1) overview of key issues concerning variable and function selection (lead by Willi Sauerbrei), (2) a primer review of spline based procedures and functions in R (lead by Aris Perperoglou), (3) review of TG2 relevant methods used in the clinical and epidemiological literature (lead by Michal Abrahamowicz). Slides of talks are available on the website and related papers are underway. Already started or considered for the next year are the following projects: (4) in depth comparison and evaluation of spline based procedures, (5) extend Heinze et al (2018)'s review on variable selection to a TG2 consensus paper about that topic. Two additional projects related to the educational part of TG2 are: (6) survey about teaching of TG2 issues to methodologists and (7) review of statistical series published in the medical literature, where we aim to support analysts with lower level of statistical knowledge (Level 1). This work could potentially have a major impact, by guiding applied researchers to avoid common misconceptions about the appropriate use of modeling strategies (Heinze & Dunkler 2017).

Longer-term goals include evaluation of and evidence based recommendations for computationally intensive variable selection algorithms, which incorporate shrinkage and resampling techniques. In this part, simulation studies, based on principles recently summarized by members of the Simulation Panel of the STRATOS initiative (Boulesteix et al 2018), will play a key role.

Several TG2 members are also active in this panel. Furthermore, the study by Binder et al. (2013), comparing the multivariable fractional polynomial (MFP) approach with some multivariable splines based procedures, needs substantial extension.

To account for Complexities such as missing data, measurement errors, time-varying confounding, or issues specific to modeling continuous predictors in survival analyses (Abrahamowicz and MacKenzie 2006) requires close collaboration with other TGs. In particular, several research topics of TG2 overlap with issues in high dimensional data, leading to a close collaboration with TG9: High-dimensional data.

References

Abrahamowicz M and MacKenzie TA (2007) Joint estimation of time-dependent and non-linear effects of continuous covariates on survival. *Stat Med* 26: 392–408.

Binder H, Sauerbrei W, Royston P (2013) Comparison between splines and fractional polynomials for multivariable model building with continuous covariates: a simulation study with continuous response. *Stat Med* **32**: 2262–2277.

Boulesteix A-L, Binder H, Abrahamowicz M, Sauerbrei W (2018) On the necessity and design of studies comparing statistical methods. *Biometrical Journal* **60**: 216–218.

Eilers PHC and Marx BD (1996) Flexible smoothing with B-splines and penalties. Statistical Science **11**: 89-102.

Harrell FE (2015) *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*, 2nd edn. Springer International Publishing; Imprint; Springer: Cham.

Hastie T and Tibshirani R (1990) *Generalized Additive Models*. Chapman & Hall/CRC: New York.

Heinze G and Dunkler D (2017) Five myths about variable selection. *Transplant* 30: 6–10.

Heinze G, Wallisch C, Dunkler D (2018) Variable selection - A review and recommendations for the practicing statistician. *Biometrical Journal* **60**: 431-449

Miller A (2002) *Subset Selection in Regression*, 2nd edn. CRC Press: Hoboken.

Royston P and Altman DG (1994) Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling. *Applied Statistics* **43**: 429–467.

Royston P and Sauerbrei W (2008) *Multivariable model-building: A pragmatic approach to regression analysis based on fractional polynomials for continuous variables*. Wiley: Chichester.

Sauerbrei W, Royston P, Binder H (2007) Selection of important variables and determination of functional form for continuous predictors in multivariable model building. *Stat Med* **26**: 5512–5528.

Wood SN (2017) *Generalized Additive Models: An Introduction with R*, 2nd edn. CRC Press: Portland.