Only fools rush in! – initial data analysis is required for developing and validating prediction models

Georg Heinze[1], Mark Baillie[2], Marianne Huebner[3]

[1]*Section for Clinical Biometrics, CeMSIIS, Medical University of Vienna; Spitalgasse 23, 1090 Vienna, Austria*
[2]*Biostatistical Sciences and Pharmacometrics, Novartis Pharma AG, Basel, Switzerland*
[3]*Department of Statistics and Probability, Michigan State University, East Lansing, MI, USA*

Background: In the age of personalized medicine, prediction models are becoming increasingly popular for risk stratification and informed treatment decisions. Accessibility of large routine data collections and observational cohorts facilitates the validation of existing prediction models and the development of new ones.
Aims: to define necessary steps and to stress the importance of initial data analysis before running regression analysis (IDA-REG), assuming that a data set has already passed an initial data cleaning stage.
Methods: Following a conceptional framework for IDA[1], we describe 3 mandatory and 3 optional steps of IDA-REG.
Results: IDA-REG focuses on informing an analyst about features in the data that should be known to the data analyst in order to a) properly interpret results of an analysis, b) make decisions on how to present the results of an analysis, and c) adapt the statistical analysis plan to avoid analysis errors. Mandatory steps include summaries of univariate distributions of predictors and outcome variable, summaries of bi- and trivariate distributions of predictors, and summaries of patterns of missing values. Optional steps include investigation of measurement error, investigation of levels of measurement (hierarchies), and exploring unsupervised possibilities to reduce dimensionality of regression models. The evaluation of associations of predictors with the outcome is explicitly not part of an IDA-REG. We exemplify IDA-REG by means of simulated and real data.
Conclusions: Appropriate graphical and analytical tools enable a researcher to perform IDA-REG in order to avoid misinterpretation, poor presentation and analysis errors. These necessary preparations are too often forgotten by inexperienced data analysts. (253 < 300 words)

Keywords
Prediction, model, data screening

References
[1] Huebner M, le Cessie S, Schmidt C, Vach W on behalf of the Topic Group “Initial Data Analysis” of the STRATOS Initiative: A Contemporary Conceptual Framework for Initial Data Analysis. Observational Studies 4 (2018):171-192.