Einladung zum Biometrischen Kolloquium

KIM LUIJKEN

Leiden University Medical Centre, The Netherlands

MEASUREMENT MATTERS: HOW DIFFERENCES IN MEASUREMENT INDUCE NON-TRANSPORTABILITY OF CLINICAL PREDICTION MODELS

25. September 2019, 9:15h

Seminarraum CeMSIIS, Raumnummer 88.03.513

Medizinischen Universität Wien, Spitalgasse 23, 1090 Wien

HOST: Georg Heinze

ABSTRACT:

Context: Transportability of prediction models can be hampered when predictors are measured differently at derivation and (external) validation. This may occur, for instance, when predictors are categorized using different cut-points or when tests are produced by different manufacturers. While such heterogeneity in predictor measurement across derivation and validation setting seems very common, little is known about the impact it may have on the performance of prediction models at external validation.

Objectives: (i) To formally define effects of predictor measurement heterogeneity on external performance of prediction models; and (ii) To illustrate the impact of predictor measurement heterogeneity in empirical data.

Methods: We examined the external performance of a predictive logistic regression model under heterogeneous predictor measurement using analytical and simulation approaches. Heterogeneity in measurement procedures was defined using a well-known taxonomy of measurement error models. Measures of predictive performance were the recalibration-model intercept and slope, c-statistic and Brier score. We additionally examined effects of various scenarios of predictor measurement heterogeneity in real-world clinical examples using previously developed prediction models for diagnosis of ovarian cancer, colorectal cancer, and intrauterine pregnancy.

Results: Our simulation study showed that heterogeneity in measurement of predictors can have a large impact on the external predictive performance of a prediction model. In empirical data, changing the measurement procedure of a predictor influenced the performance at validation of the prediction models in nine clinical examples. Notably, it induced model miscalibration. The calibration-in-the-large coefficient at validation ranged from -0.70 to 1.43 (0 for good calibration), while the calibration slope ranged from 0.50 to 1.67 (1 for good calibration). The difference in c-statistic and scaled Brier score between derivation and validation ranged from -0.08 to +0.08 and from -0.40 to +0.16, respectively.

Conclusions: Our work highlights measurement heterogeneity as an important explanation of unanticipated (good or bad) out-of-sample performance of a clinical prediction model. We recommend that prediction models are derived from and validated on datasets collected with measurement procedures that are in widespread use in the intended clinical setting.