Propensity-based standardization to enhance the interpretation of predictive performance in individual participant data meta-analysis

Valentijn M.T. de Jong[1], Jeroen Hoogland[1], Karel G.M. Moons[1,2], Tri-Long Nguyen[1,3,4], Thomas P.A. Debray[1,2]

[1]*Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands*
[2] *Cochrane Netherlands, University Medical Center Utrecht, Utrecht, The Netherlands*
[3] *Section of Epidemiology, Department of Public Health, University of Copenhagen, Copenhagen, Denmark*
[4] *Department of Pharmacy, University Hospital Centre of Nîmes and University of Montpellier, Nîmes, France*

Background: Meta-analysis of individual participant data (IPD-MA) offers new opportunities for studying the generalizability of prediction models across different settings and populations. The interpretation of model performance estimates in IPD-MA is often challenging, because between-study heterogeneity may arise from invalid model coefficients and differences in (the distribution of) population characteristics. Hence, the benefit of local model revisions may be unclear.

Aim: To disentangle the effects of differences in case-mix and invalid regression coefficients, to allow for the identification of reproducibility of model performance and predictor effects.

Methods: We propose to standardize the c-statistic, calibration slope and calibration-in-the-large for case-mix differences between samples by applying propensity-weighting. The propensity scores are derived using a (multinomial) membership model that predicts the originating sample of an individual in the IPD-MA. We illustrate our methods in a motivating example on the validation of eight diagnostic prediction models for detecting deep vein thrombosis (DVT) that may aid in the diagnosis of patients suspected of DVT in 12 external validation data sets. We analyze the estimates of prediction models' performance across the external validation sets with random effects meta-analysis.

Results: In the meta-analysis of c-statistics, summary estimates were not affected much by standardization. However, standardization substantially reduced the between-study heterogeneity, indicating that variation of the models' discrimination across the validation studies can partially be attributed to differences in case-mix, rather than invalid model coefficients.
Standardization increased the estimated between-study heterogeneity in calibration slopes. This implies that the predictor effects do not reproduce well in new samples with the same case-mix distribution.

Conclusion: Propensity score-based standardization may facilitate the interpretation of (heterogeneity in) prediction model performance across external validation studies, guide model updating strategies or show that the validation sample does not reflect the target population of the model.