Large-scale validation of the Prediction model Risk Of Bias ASsessment Tool (PROBAST) using a short form

Esmee Venema[1,2], Benjamin S Wessler[3,4], Jessica K Paulus[3], Rehab Salah[5], Gowri Raman[6], Lester Y Leung[7], Benjamin C Koethe[3], Jason Nelson[3], Jinny G Park[3], David van Klaveren[1,3], Ewout W Steyerberg[1,8], and David M Kent[3]

[1]Department of Public Health, Erasmus MC University Medical Center, Rotterdam, the Netherlands; [2]Department of Neurology, Erasmus MC University Medical Center, Rotterdam, the Netherlands; [3]Predictive Analytics and Comparative Effectiveness Center, Tufts Medical Center, Boston, MA, USA; [4]Division of Cardiology, Tufts Medical Center, Boston, MA, USA; [5]Benha Faculty of Medicine, Benha, Egypt; [6]Center for Clinical Evidence Synthesis, Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, MA, USA; [7]Division of Stroke and Cerebrovascular Diseases, Department of Neurology, Tufts Medical Center, Boston, MA, USA; [8]Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the Netherlands.

Background: The comprehensive Prediction model Risk Of Bias ASsessment Tool (PROBAST)[1] was developed for reviews of clinical prediction models (CPMs).

Aims: To assess whether PROBAST can identify CPMs that perform poorly at external validation and to develop a short form that is equally capable to identify poorly performing CPMs.

Methods: We evaluated risk of bias (ROB) using the PROBAST on 102 CPMs from the Tufts Predictive Analytics and Comparative Effectiveness Registry, compared to a short form consisting of six of the 20 PROBAST items anticipated to best identify high ROB. We then applied the short form to all CPMs in the Registry with at least 1 validation (n=556). Primary outcome was the change in the area under the receiver operating characteristic curve (dAUC, available for 1,147 validations) between the derivation and the validation cohorts in low versus high ROB CPMs.

Results: The full PROBAST classified 98 of 102 CPMS as high ROB. The short form identified 96 of these 98 as high ROB (98% sensitivity), with perfect specificity. Perfect agreement with the full PROBAST could be achieved with re-review of only a small number of low ROB CPMs. In the full CPM registry, 529 of 556 CPMs (95%) were classified as high ROB, 20 (4%) low ROB, and 7 (1%) unclear ROB. The median change in discrimination was significantly smaller in low ROB models (dAUC -0.9%, IQR -6.2% to 4.2%) compared to high ROB models (dAUC -12%, IQR -33% to 2.6%; p<0.001).

Conclusions: High ROB is pervasive among published CPMs. It is associated with poor performance at validation, supporting the application of PROBAST or a shorter version in reviews of CPMs.

Keywords
Prediction models, risk of bias, evaluation

References
[1]R.F. Wolff, K.G.M. Moons, R.D. Riley, et al., *Ann Intern Med*, 170 **2019**, 51-58.