Journal of Clinical Epidemiology 138 (2021) 32–39

ORIGINAL ARTICLE

Large-scale validation of the prediction model risk of bias assessment Tool (PROBAST) using a short form: high risk of bias models show poorer discrimination

Esmee Venema[a,b], Benjamin S. Wessler[c,d], Jessica K. Paulus[c], Rehab Salah[e], Gowri Raman[f], Lester Y. Leung[g], Benjamin C. Koethe[c], Jason Nelson[c], Jinny G. Park[c], David van Klaveren[a,c], Ewout W. Steyerberg[a,h], David M. Kent[c,*]

[a] *Department of Public Health, Erasmus MC University Medical Center, Rotterdam, the Netherlands*
[b] *Department of Neurology, Erasmus MC University Medical Center, Rotterdam, the Netherlands*
[c] *Predictive Analytics and Comparative Effectiveness Center, Tufts Medical Center, Boston, MA, USA*
[d] *Valve Center, Division of Cardiology, Tufts Medical Center, Boston, MA, USA*
[e] *Ministry of Health and Population Hospitals, Benha Faculty of Medicine, Benha, Egypt*
[f] *Center for Clinical Evidence Synthesis, Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, MA, USA*
[g] *Comprehensive Stroke Center, Division of Stroke and Cerebrovascular Diseases, Department of Neurology, Tufts Medical Center, Boston, MA, USA*
[h] *Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the Netherlands*

Accepted 21 June 2021; Available online 24 June 2021

Abstract

Objective: To assess whether the Prediction model Risk Of Bias ASsessment Tool (PROBAST) and a shorter version of this tool can identify clinical prediction models (CPMs) that perform poorly at external validation.

Study Design and Setting: We evaluated risk of bias (ROB) on 102 CPMs from the Tufts CPM Registry, comparing PROBAST to a short form consisting of six 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) and assessed the change in discrimination (dAUC) in external validation cohorts (n=1,147).

Results: PROBAST classified 98/102 CPMS as high ROB. The short form identified 96 of these 98 as high ROB (98% sensitivity), with perfect specificity. In the full CPM registry, 527 of 556 CPMs (95%) were classified as high ROB, 20 (3.6%) low ROB, and 9 (1.6%) unclear ROB. Only one model with unclear ROB was reclassified to high ROB after full PROBAST assessment of all low and unclear ROB models. Median change in discrimination was significantly smaller in low ROB models (dAUC -0.9%, IQR -6.2–4.2%) compared to high ROB models (dAUC -11.7%, IQR -33.3–2.6%; $P<0.001$).

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

Keywords: Prognosis; Prediction model; Validation; Bias; Risk; Cardiovascular disease

Funding: This research was funded by a Patient-Centered Outcomes Research Institute (PCORI) Methods Award (ME-1606-35555). The authors declare that the funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

* Corresponding author: D.M.K
E-mail address: dkent1@tuftsmedicalcenter.org (D.M. Kent).

https://doi.org/10.1016/j.jclinepi.2021.06.017

What is new?

- High risk of bias is pervasive among published clinical prediction models.
- High risk of bias identified with PROBAST is associated with poorer discriminative performance at validation.
- A subset of questions can distinguish between models with high and low risk of bias.

1. Introduction

A clinical prediction model (CPM) estimates an individual's probability of having a disease (diagnosis) or developing a clinical outcome (prognosis) based on the combination of relevant characteristics. Such models enable researchers or clinicians to inform individuals about their expected outcome and can be used to select patients for a certain treatment or study [1–3]. The development of a CPM consists of several important steps, including careful predictor selection and model specification [4,5]. Methodological shortcomings may cause bias and systematic overestimation of the model performance measures, promoting misleading conclusions of its value for clinical practice. External validation in a population other than the original derivation cohort is important for CPMs to assess their generalizability and transportability [2,6,7].

The Prediction model Risk Of Bias ASsessment Tool (PROBAST) was developed to assess risk of bias (ROB) based on the methods used for model development [8]. It was developed for systematic reviews and provides a comprehensive overview of methodological quality. It is unclear whether adherence to the methodological standards of PROBAST is associated with a better performance in external validation. Moreover, PROBAST requires both subject and methodological expertise to apply and might be too time intensive for large-scale use. We aimed to examine whether poor PROBAST scores can identify CPMs that perform poorly at external validation and to develop a short form that is equally capable to identify poorly performing CPMs.

2. Methods

We used publications from the Tufts Predictive Analytics and Comparative Effectiveness (PACE) CPM Registry, a database which includes CPMs for cardiovascular disease published in English language from January 1990 through March 2015 (www.pacecpmregistry.org). A systematic literature search was performed to comprehensively identify CPMs [9,10]. For this registry, a de novo CPM was defined as a newly derived prediction model to estimate an individual patient's absolute risk for a binary outcome. Information from each CPM was extracted from the original article and entered in the database. CPMs were characterized based on the index condition of the patients to whom the model applies, including coronary artery disease, congestive heart failure, arrhythmias, stroke, venous thromboembolism, and peripheral vascular disease. Populations of healthy individuals at risk for developing incident CVD were classified with the index condition of 'population sample'.

2.1. External validations

We included validations of each de novo CPM in the CPM Registry, which have been previously identified with a Scopus citation search conducted on March 22, 2017 [11]. An external validation was defined as any model evaluation on a dataset distinct from the derivation data, including validations that were performed on a temporally or geographically distinct part of the same cohort (i.e., nonrandom split sample), that reported at least one measure of model performance (discrimination and/or calibration). Discrimination, expressed as the area under the receiver operating characteristic curve (AUC), describes how well a model separates those who develop the outcome of interest from those who do not. Calibration refers to the agreement between the observed and predicted probabilities. Due to the universality of the AUC as a metric for model performance, we used the difference in discriminative ability between the derivation and validation cohort (dAUC) as the endpoint in this study. Although calibration measures should preferably be evaluated as well, this was limited by the large variation in and underreporting of calibration measures in the published articles.

To assess clinical relatedness of the derivation and validation populations, relatedness rubrics were constructed for the 10 most common index conditions and included details on the population sample, major inclusion and exclusion criteria, outcome measure, enrollment period, type of intervention, and follow-up duration. An example is provided in the Supplemental Material (Table S1) and further details of these rubrics are available elsewhere [11]. Items were extracted from the CPM Registry and, if necessary, the original articles. The relatedness rubrics were used to distinguish between 'related' populations, with a near exact match on relevant items for each specific index condition, and 'distantly related' populations. Validation cohorts with a different index condition were excluded from analysis.

2.2. PROBAST assessment

The PROBAST tool assesses risk of bias based on 20 signalling questions in 4 key domains: participants (e.g., study design and patient inclusion), predictors (e.g., differences in predictor definitions), outcome (e.g., differences in outcome assessment), and analyses (e.g., sample size and handling of missing data). All questions are phrased so that "yes" indicates low ROB and "no" high ROB. As

a result, a domain where all questions are answered with "yes" is rated low ROB, while questions answered with "no" indicate a high ROB. Unclear ROB is assigned to a domain when there is insufficient information to assess one or more questions. The overall judgment is considered low ROB when all domains have low ROB. If at least 1 domain has high ROB, the overall judgment is high ROB as well. If at least 1 domain has unclear ROB and all other domains have low ROB, the overall judgment is unclear ROB [8,12].

We applied the PROBAST tool to all stroke models with at least one external validation and a reported derivation AUC (n=52) and 50 randomly selected models with other index conditions. Each of these 102 models was assessed by two independent reviewers—blinded to validation study results–using the guidelines in the PROBAST 'Explanation and Elaboration' [12]. Discrepancies were discussed with a third reviewer to arrive at a consensus.

2.3. Short form

For practical reasons, a selection was made from items in the original PROBAST. We discussed the relevance of all 20 questions within a group of methodologists with specialized expertise in prediction (including DvK, EWS, and DMK). We rated items according to their potential effect on performance of a CPM at validation. We aimed for a subset of questions that can be applied by a trained research assistant (i.e., trained study staff without doctoral-level clinical or statistical expertise) within 15 minutes. Usability of the items was verified through testing by experienced research assistants at Tufts PACE Center.

The following 6 items were considered most relevant and easy to use: 'outcome assessment', 'events per variable (EPV)', 'continuous predictors', 'missing data', 'univariable analysis', and 'correction for overfitting/optimism'. We adjusted the definitions from the original PROBAST article to improve clarity and resolve unambiguity (see Supplemental Material). One point was assigned for each item that was incorrectly performed, resulting in total scores ranging from 0 to 6 points. As with the full PROBAST, models with a total score of 0 were classified as 'low ROB' and models with a score $\geq$1 as 'high ROB'. When the total score was 0 but there was insufficient information provided to assess all items, the model was rated 'unclear ROB'. We assumed that the effect of using univariable selection or not correcting for optimism would be negligible when the effective sample size was very large. Hence, we did not assign points to these items when the EPV was $\geq$25 for candidate predictors, or $\geq$50 for the final model (when candidate predictors were unknown).

We applied this short form to the same set of 102 models and compared results with those of the full PROBAST. We then refined and clarified the scoring guidelines. Research assistants of Tufts PACE center then applied the short form to all de novo CPMs with at least one external validation in the Registry (n=556). Blinded double assessment of the first 40 models was done to compare the assessments and discuss discrepancies. Because the short form was composed of a subset of PROBAST items, CPMs classified as high ROB by the short form were anticipated to also be classified as high ROB by the full PROBAST. CPMs classified by the short form as low ROB might be reclassified as high ROB by the full PROBAST. Thus, all models that were rated as low or unclear ROB were reassessed by a separate reviewer using PROBAST to reveal any potential items suggestive of high ROB not captured by the short form.

2.4. Analyses

Cohen's kappa statistic was calculated to assess interrater reliability and agreement between PROBAST and the short form. As a measure of the observed ROB in each derivation-validation pair, we used the change in discriminative performance between the derivation and validation cohorts, as quantified by the AUC. The AUC ranges from 0.5 (similar to a coin flip) to 1.0 (perfect discrimination) [13]. The percent change in discrimination was thus calculated:

$$\%\ change\ in\ discrimination = \frac{(validation\ AUC - 0.5) - (derivation\ AUC - 0.5)}{(derivation\ AUC - 0.5)}\ x\ 100$$

For example, when the AUC decreases from 0.70 in derivation to 0.60 in validation, the delta AUC (dAUC) represents a 50% loss in discriminative ability (since 0.50 is the reference value for AUC). We calculated the median and interquartile range (IQR) of the change in discrimination for low ROB versus high ROB models and stratified for relatedness.

We used generalized estimation equations (GEE) with robust covariance estimator [14,15] to assess the association between the ROB classification and the observed change in discrimination, taking into account the correlation between validations of the same CPM. We used the binomial distribution and a logit link function for the GEE model. The estimated outcome in these analyses was the absolute difference between two dAUCs. For example, if one model had dAUC 20%, while the other model had a dAUC 6%, the difference in dAUC would be 14%. We also constructed a multivariable GEE model to control for the following factors: relatedness, index condition, CPM authors (same as in validation paper, author overlap, no author overlap), CPM method (logistic, time-to-event, other), CPM center (single, multicenter), CPM source (medical record, registry, trial, other), validation design (cohort, trial, other), validation center (single, multicenter), validation source (medical record, registry, trial, other), CPM parameter degrees of freedom, CPM events per variable, CPM sample size, CPM events, validation events per variable,

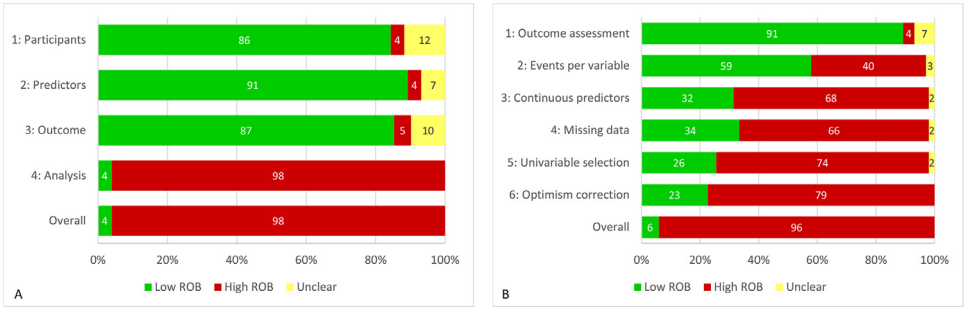

Fig. 1. Risk of bias assessed with PROBAST (A, per domain) and the short form (B, per item) in the initial set of 102 clinical prediction models. The first item of the short form ("Outcome assessment") belongs to domain 3 ("Outcome"), all other items belong to domain 4 ("Analysis"). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

validation sample size, validation events, relative outcome rate difference >40%, and difference in years between CPM and validation. For this multivariable model, multiple imputation using 20 imputation sets was used to account for missing data.

All statistical analyses were performed using SAS Enterprise Guide version 8.2 (SAS Institute Inc., Cary, NC, USA).

3. Results

PROBAST was assessed on the first set of 102 models (52 stroke models and 50 models with other index conditions). Of these models, 98 (96%) were rated high ROB and only 4 (3.9%) low ROB. Overall high ROB was mainly caused by high ROB in the analysis domain, while the other three domains contributed little information (Fig. 1A). Agreement between the two reviewers before the final consensus meeting was 90% for the overall judgment (kappa 0.33). Interrater agreement ranged between 49 and 97% per item (kappa -0.05 to 0.88, Supplemental Table S2). When applying the short form to the same 102 models, the sensitivity to detect high ROB was 98% (using the full PROBAST as reference standard) and specificity was 100%. Overall agreement was good (98%, kappa 0.79). The item 'outcome assessment' was rated high ROB in only 4% of the models, while the percentage high ROB of the other items ranged between 39 and 77% (Fig. 1B). Fig. 2 shows the distribution of the short form total scores.

The CPM Registry included 1,382 de novo CPMs, of which 556 (40%) were externally validated at least once. The most common index conditions were coronary artery disease (n=129), stroke (n=97), population sample (n=85), and cardiac surgery (n=70). The reported number of events per variable in the final model ranged between 1.5 and 4,858 with a median EPV of 26 (IQR 12 – 71). In total, 1,846 validations of the 556 CPMs were included in the CPM Registry. The number of validations per model ranged from 1 to 86 with a median of 1 (IQR 1 – 3). Relatedness was assessed for 1,702 validations (i.e., of CPMs for the top 10 index conditions), of which 985

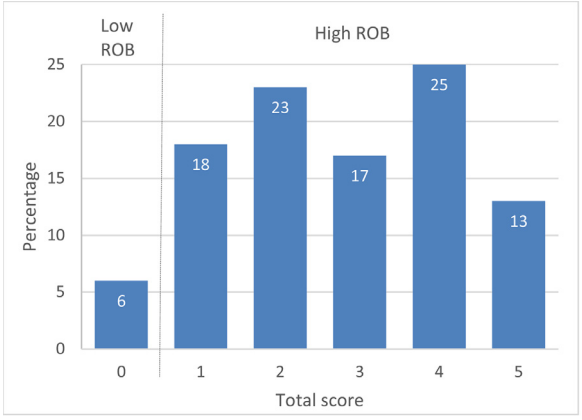

Fig. 2. Distribution of the short form total scores in the initial set of 102 clinical prediction models. Total score can range from 0 to 6, with 0 points indicating low risk of bias and ≥1 high risk of bias. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

(58%) were related and 717 (42%) were distantly related. Following further clarifications of guidance (see Supplemental Material), the short form was applied to all 556 de novo models in the dataset. In total, 527 (95%) were considered high ROB, 20 (3.6%) low ROB, and 9 (1.6%) unclear ROB. Only one model with unclear ROB was reclassified to high ROB after full PROBAST assessment of all low and unclear ROB models.

Information on both the derivation AUC and validation AUC was available for 1,147 validations (62%). The median dAUC of all derivation-validation pairs was -10.7% (IQR -32.4% – 2.8%). The difference was significantly smaller in low ROB models (dAUC -0.9%; IQR -6.2% – 4.2%) compared to high ROB models (dAUC -11.7%; IQR -33.3% – 2.6%; P <.001; Table 1 and Fig. 3). Including related validations only (n=623), the median dAUC was -8.8% (IQR -26.2% – 3.8%) in high ROB models and -1.4% (IQR -6.2% – 3.6%) in low ROB models. While our main analysis treated ROB as a binary measure, a "dose-response" relationship was observed among high ROB studies; the median dAUC was generally lower

Table 1. Percent change in discrimination between derivation AUC and validation AUC (dAUC)

	N	N missing	N included	Median dAUC (IQR)	Range
All validations	1846	699	1147	-10.7% (-32.4% – 2.8%)	-179% – 191%
High ROB	1780	667	1113	-11.7% (-33.3% – 2.6%)	-179% – 191%
Low ROB	39	11	28	-0.9% (-6.2% – 4.2%)	-20% – 45%
Unclear ROB	27	21	6	-13.3% (-25.7% – -4.1%)	-42% – 0%
Related validations	985	362	623	-8.3% (-25.2% – 3.8%)	-129% – 157%
High ROB	947	348	599	-8.8% (-26.2% – 3.8%)	-129% – 157%
Low ROB	28	8	20	-1.4% (-6.2% – 3.6%)	-20% – 11%
Unclear ROB	10	6	4	-8.6% (-13.3% – -2.1%)	-14% – 0%
Distantly related validations	717	286	431	-17.2% (-42.1% – 0.2%)	-133% – 191%
High ROB	692	269	423	-17.4% (-42.4% – 0%)	-133% – 191%
Low ROB	8	2	6	4.2% (3.3% – 27.3%)	-12% – 31%
Unclear ROB	17	15	2	-33.9% (-42.1% – -25.7%)	-42% – -26%

Relatedness was assessed using relatedness rubrics specifically developed for this purpose [11].
AUC indicates area under the receiver operator characteristic curve; IQR, interquartile range; ROB, risk of bias.

in models with fewer points on the short form (Supplemental Table S3). Low ROB studies also reported calibration more often than high ROB studies (calibration plot: 50% vs. 19%, any calibration measure (including goodness-of-fit tests): 75% vs. 47%).

The GEE analyses estimated a difference in dAUC of 16.8% (95% CI 6.1% to 27.6%; $P = 0.002$) for low ROB versus high ROB after adjustment for CPM and validation characteristics (Table 2). This number can be interpreted as an absolute difference between the validation AUC of a high ROB model and a low ROB model of 0.02 when the derivation AUC was 0.60 or as 0.07 when the derivation AUC was 0.90 (Supplemental Table S4). Other parameters in the multivariable model with a statistically significant difference in dAUC were the validation center (single vs. multicenter: dAUC 9.4%; 95% CI 4.2% – 14.7%; $P < .001$) and years between development and validation (each year: dAUC -1.2%; 95% CI -1.8% to -0.5%; $P < .001$; Supplemental Table S5).

4. Discussion

PROBAST can be used to obtain a comprehensive overview of important methodological elements of CPM studies. A selection of 6 key items from the original 20 items fully captured the overall ROB assessment. We assessed a large set of CPMs of which most models were classified as high ROB, which was associated with poorer discriminative ability at external validation compared to CPMs with low ROB scores.

While we did not formally conduct time assessments, applying the full PROBAST took up to one hour and required both subject and methodological expertise. While some items may be more likely to cause or identify optimism in model performance than others, the tool does not prioritize items. For example, a key pitfall ofusing a small dataset with an inadequate number of events-per-variable is that it is weighted the same as not reporting all relevant performance measures. Thus, in constructing a short form, we prioritized those items felt to be most relevant for poor performance at validation. Moreover, while the 'Explanation and Elaboration' article provides extensive guidelines for the assessment [12], it nevertheless leaves some questions open for interpretation, complicating assessment by research assistants without specific training. We clarified guidelines to diminish ambiguity and reduce interrater variability. The average time of applying the selected subset of 6 questions was approximately 15 minutes.

The selection of items for the short form, which was done strictly by expert opinion and prioritized items in the analysis domain, showed concurrent validity with the PROBAST assessment in 102 papers. Nearly all high ROB CPMs were identified based on violating any item from the analysis domain, which was well represented in the short

Table 2. Results of generalized estimated equations (GEE) performed with data of 1048 validations

	dAUC (95% CI)		Difference in dAUC (95% CI)	*P*-value
Unadjusted analysis				
Related	-10.1% (-13.3% to 6.8%)	Related vs. distantly related*	9.7% (3.8% to 15.7%)	<0.001
Distantly related	-19.8% (-25.7% to -13.9%)			
Low ROB	0.8% (-4.5% to 6.0%)	Low ROB vs. high ROB†	14.7% (8.5% to 21.0%)	<0.001
High ROB	-14.0% (-17.3% to -10.6%)			
Adjusted analysis‡				
Related	-4.1% (-8.0% to 0.2%)	Related vs. distantly related	9.6% (3.6% – 15.5%)	0.002
Distantly related	-13.7% (-20.0% to -7.3%)			
Low ROB	-2.3% (-9.6% – 5.0%)	Low ROB vs. high ROB	13.2% (5.4% – 20.9%)	<0.001
High ROB	-15.5% (-19.3% – -11.7%)			
Multivariable model§				
		Related vs distantly related	8.3% (2.2% to 14.4%)	0.007
		Low ROB vs high ROB	16.8% (6.1% to 27.6%)	0.002

All models exclude validations not assessed for relatedness (n=144), dAUC is missing (n=648), or risk of bias is "unclear" (n=6). Relatedness was assessed using relatedness rubrics specifically developed for this purpose. [11]

dAUC indicates the change in area under the receiver operator characteristic curve; CI, confidence interval; CPM, clinical prediction model; ROB, risk of bias.

* When including "unclear" risk of bias (n=1054): difference 9.8% (95% CI 3.9% to 15.8%); P=.001.

† When including missing relatedness assessment (n=1141): difference 16.2% (95% CI 9.9% to 22.5%); P<.001.

‡ Model includes relatedness and risk of bias. Interaction between relatedness and risk of bias was tested in a separate model: interaction p-value = 0.17.

§ Model (20 imputed data sets) includes: relatedness, risk of bias, index condition, CPM authors (same as in validation paper, author overlap, no author overlap), CPM method (logistic, time-to-event, other; missing = 20), CPM center (single, multicenter; missing = 23), CPM source (medical record, registry, trial, other; missing = 4), validation design (cohort, trial, other; missing = 18), validation center (single, multicentre; missing = 98), validation source (medical record, registry, trial, other; missing = 44), CPM parameter degrees of freedom, CPM events per variable (missing = 208), CPM sample size (missing = 29), CPM events (missing = 208), validation events per variable (missing = 147), validation sample size (missing = 6), validation events (missing = 147), relative outcome rate difference >40% (missing = 372), and difference in years between CPM and validation.

form. Additionally, because the short form is comprised of a subset of the PROBAST questions, it is 100% specific for identifying low ROB CPMs. Because the number of low ROB CPMs is very small (typically <5% of the total), review of this subset with the full PROBAST should result in identical classification as using PROBAST for all CPMs while taking a fraction of the time and expertise. The original PROBAST can only reclassify some low ROB CPMs (according to the short form) to high ROB, not the other way around. In this particular sample, only a single CPM was so reclassified from low to high ROB.

Most CPMs in our study were rated high ROB, in line with previous systematic reviews using PROBAST [16–20]. This is inherent to the structure of PROBAST: one incorrectly performed item in one domain determines an overall judgment of high ROB. While the high percentage of models with high ROB might be interpreted as a reflection of the low overall quality of the literature, it might also reflect limitations of the tool. We found substantial variation in the number of items violated by any individual CPMs, which suggests variation in methodological rigor among high ROB CPMs. Moreover, the discriminatory performance within the high ROB group varied widely (IQR of dAUC ranging from -33% to +2.6%). While the purpose of this study was to validate the ROB assessment by PROBAST, future work might explore whether it is useful to identify an "intermediate" ROB category, which might provide a more graded, less stringent assessment. Indeed, our preliminary analysis suggests just such a dose-response relationship.

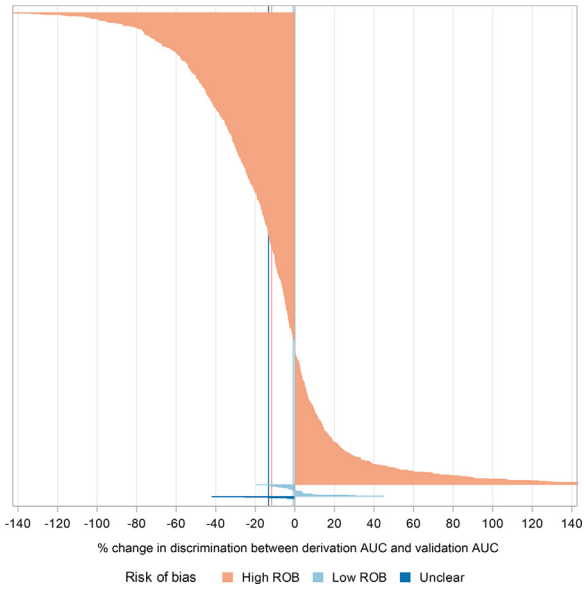

Fig. 3. Waterfall plot showing the distribution of dAUC in each derivation-validation pair (n=1,147), divided by the risk of bias. AUC indicates area under the receiver operating characteristic curve; ROB, risk of bias. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Our study has several limitations. Poor model performance at external validation can be due to the relatedness of the settings where the CPM is developed versus validated, changes in case-mix, and methodological problems (i.e., statistical overfitting) [21,22]. Our analyses were focused on ROB caused by methodological issues, while relatedness of the validation cohort and case-mix differences would also affect the observed change in discrimination [23,24]. Relatedness rubrics were used to adjust for differences in relatedness of the validation cohort. The effect of case mix differences is potentially measurable through a model based concordance (c) statistic [25,26] but this requires primary patient-level data to compute. Furthermore, discrimination is not the only, or even necessarily the most important, metric by which to evaluate model performance. The net benefit of applying a model for decision support in a new population is a function of both discrimination and calibration [27]. Calibration might be more sensitive to bias in model development, since it depends on the consistency of both measured and unmeasured predictor effects. However, calibration metrics are incompletely and inconsistently reported; when reported, the metrics provided are usually either clinically uninformative (e.g., the Hosmer-Lemeshow test) or difficult to quantitatively analyze (e.g., graphical) [21]. We also found that relevant information on model development was often lacking, for example about the selection of predictors, or handling of missing data. This emphasizes the importance of standardizing the reporting of CPM studies by conforming to the TRIPOD guidelines [28,29]. Recent work on sample size has shown that heuristics such as EPV can be misleading, so our thresholds for adequate sample size may not be wholly appropriate for all CPMs [30,31]. Also, it is possible that our search did not comprehensively identify all external validations, although the number of inadvertently excluded studies should be small.

Personalized, risk-based decision making has become increasingly important in medicine, leading to a substantial increase in the number of published CPMs over the past 2 decades [2]. Many prediction models are available, though their use in clinical practice is often limited [32–34]. Also, the quality of reporting is considered variable and insufficient [35–39]. Therefore, reviews of CPM studies are important to help clinicians and policy makers decide which CPM should be promoted in evidence-based guidelines or implemented in practice. The PROBAST scores, either from the full instrument or from a subset of questions, appear valid to assess and compare model quality. A recent systematic review of prediction models for COVID-19 identified 145 models that were all high ROB, mostly based on a combination of poor reporting and poor methodological conduct [40]. Adequate sample size and a rigorous methodological approach are required to develop robust CPMs that can be applied in clinical practice.

In conclusion, high ROB is pervasive and is associated with poorer discriminative performance at validation, supporting the application of PROBAST in reviews of CPMs. A subset of questions from PROBAST may be particularly useful for high volume assessments, when classification into high and low ROB categories is the primary goal. Furthermore, the high prevalence of high ROB models emphasizes the need to improve methodological quality of prediction research.

Acknowledgements

We thank Christine M. Lundquist, Rebecca E. H. Maunder, and Jennifer S. Lutz (research assistants at the Predictive Analytics and Comparative Effectiveness (PACE) Center, Tufts Medical Center, Boston, MA) for their help with the data extraction and assessment of the short form.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jclinepi.2021.06.017.

References

[1] Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? BMJ 2009;338:b375.
[2] Steyerberg EW. Clincial. Prediction models: a practical approach to development, validation and updating. 2nd ed. New York: Springer; 2019.
[3] Yu T, Vollenweider D, Varadhan R, Li T, Boyd C, Puhan MA. Support of personalized medicine through risk-stratified treatment

recommendations - an environmental scan of clinical practice guidelines. BMC Med 2013;11:7.
[4] Royston P, Moons KG, Altman DG, Vergouwe Y. Prognosis and prognostic research: developing a prognostic model. BMJ 2009;338:b604.
[5] Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J 2014;35(29):1925–31.
[6] Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. BMJ 2009;338:b605.
[7] Steyerberg EW, Nieboer D, Debray TPA, van Houwelingen HC. Assessment of heterogeneity in an individual participant data meta–analysis of prediction models: an overview and illustration. Stat Med 2019;38(22):4290–309.
[8] Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019;170(1):51–8.
[9] Wessler BS, Lai Yh L, Kramer W, Cangelosi M, Raman G, Lutz JS, et al. Clinical prediction models for cardiovascular disease: tufts predictive analytics and comparative effectiveness clinical prediction model database. Circ Cardiovasc Qual Outcomes 2015;8(4):368–75.
[10] Wessler BS, Paulus J, Lundquist CM, Ajlan M, Natto Z, Janes WA, et al. Tufts PACE clinical predictive model registry: update 1990 through 2015. Diagn Progn Res 2017;1:20.
[11] Wessler BS, Nelson J, Park JG, McGinnes H, Gulati G, Brazil R, et al. External validations of cardiovascular clinical prediction models: a large-scale review of the literature. Circ Cardiovasc Qual Outcomes 2021 in press.
[12] Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med 2019;170(1):W1–W33.
[13] Harrell FJ Jr. Regression modeling strategies. New York: Springer; 2001.
[14] Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. Biometrics 1986;42(1):121–30.
[15] Zeger SL, Liang KY, Albert PS. Models for longitudinal data: a generalized estimating equation approach. Biometrics 1988;44(4):1049–60.
[16] Bellou V, Belbasis L, Konstantinidis AK, Tzoulaki I, Evangelou E. Prognostic models for outcome prediction in patients with chronic obstructive pulmonary disease: systematic review and critical appraisal. BMJ 2019;367:l5358.
[17] Aladwani M, Lophatananon A, Ollier W, Muir K. Prediction models for prostate cancer to be used in the primary care setting: a systematic review. BMJ Open 2020;10(7):e034661.
[18] Brown FS, Glasmacher SA, Kearns PKA, MacDougall N, Hunt D, Connick P, et al. Systematic review of prediction models in relapsing remitting multiple sclerosis. PLoS One 2020;15(5):e0233575.
[19] Cooray SD, Wijeyaratne LA, Soldatos G, Allotey J, Boyle JA, Teede HJ. The unrealized potential for predicting pregnancy complications in women with gestational diabetes: a systematic review and critical appraisal. Int J Environ Res Public Health 2020;17(9).
[20] Di Tanna GL, Wirtz H, Burrows KL, Globe G. Evaluating risk prediction models for adults with heart failure: A systematic literature review. PLoS One 2020;15(1):e0224135.
[21] Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology (Cambridge, Mass) 2010;21(1):128–38.
[22] Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med 1999;130(6):515–24.
[23] Vergouwe Y, Moons KG, Steyerberg EW. External validity of risk models: use of benchmark values to disentangle a case-mix effect from incorrect coefficients. Am J Epidemiol 2010;172(8):971–80.
[24] Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol 2015;68(3):279–89.
[25] van Klaveren D, Gönen M, Steyerberg EW, Vergouwe Y. A new concordance measure for risk prediction models in external validation settings. Stat Med 2016;35(23):4136–52.
[26] van Klaveren D, Steyerberg EW, Gonen M, Vergouwe Y. The calibrated model-based concordance improved assessment of discriminative ability in patient clusters of limited sample size. Diagn Progn Res 2019;3:11.
[27] Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ (Clinical research ed) 2016;352:i6.
[28] Hemingway H, Croft P, Perel P, Hayden JA, Abrams K, Timmis A, et al. Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes. BMJ 2013;346:e5595.
[29] Collins GS, Reitsma JB, Altman DG, Moons KG, Group T. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. The TRIPOD Group. Circulation 2015;131(2):211–19.
[30] Riley RD, Snell KI, Ensor J, Burke, DL, Harrell FE Jr, Moons KG, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. Stat Med 2019;38(7):1276–96.
[31] Snell KIE, Archer L, Ensor J, Bonnett LJ, Debray TPA, Phillips B, et al. External validation of clinical prediction models: simulation-based sample size calculations were more reliable than rules-of-thumb. J Clin Epidemiol 2021;135:79–89.
[32] Kappen TH, van Klei WA, van Wolfswinkel L, Kalkman CJ, Vergouwe Y, Moons KGM. Evaluating the impact of prediction models: lessons learned, challenges, and recommendations. Diagn Progn Res 2018;2:11.
[33] Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ (Clinical research ed) 2009;338:b606.
[34] Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. Ann Intern Med 2006;144(3):201–9.
[35] Mallett S, Royston P, Dutton S, Waters R, Altman DG. Reporting methods in studies developing prognostic models in cancer: a review. BMC Med 2010;8:20.
[36] Collins GS, Mallett S, Omar O, Yu LM. Developing risk prediction models for type II diabetes: a systematic review of methodology and reporting. BMC Med 2011;9:103.
[37] Bouwmeester W, Zuithoff NP, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, et al. Reporting and methods in clinical prediction research: a systematic review. PLoS Med 2012;9(5):1–12.
[38] Collins GS, Omar O, Shanyinde M, Yu LM. A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods. J Clin Epidemiol 2013;66(3):268–77.
[39] Siontis GC, Tzoulaki I, Castaldi PJ, Ioannidis JP. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. J Clin Epidemiol 2015;68(1):25–34.
[40] Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. BMJ 2020;369:m1328.