



Universiteit
Leiden
The Netherlands

Appraising prediction research: a guide and meta-review on bias and applicability assessment using the Prediction model Risk Of Bias ASsessment Tool (PROBAST)

Jong, Y. de; Ramspek, C.L.; Zoccali, C.; Jager, K.J.; Dekker, F.W.; Diepen, M. van

Citation

Jong, Y. de, Ramspek, C. L., Zoccali, C., Jager, K. J., Dekker, F. W., & Diepen, M. van. (2021). Appraising prediction research: a guide and meta-review on bias and applicability assessment using the Prediction model Risk Of Bias ASsessment Tool (PROBAST). *Nephrology*, 26(12), 939-947. doi:10.1111/nep.13913


Version: Publisher's Version

License: [Creative Commons CC BY-NC 4.0 license](https://creativecommons.org/licenses/by-nc/4.0/)

Downloaded from: <https://hdl.handle.net/1887/3276154>

Note: To cite this publication please use the final published version (if applicable).

Appraising prediction research: a guide and meta-review on bias and applicability assessment using the Prediction model Risk Of Bias ASsessment Tool (PROBAST)

Ype de Jong^{1,2}  | Chava L. Ramspek¹ | Carmine Zoccali^{3,4} | Kitty J. Jager⁵ | Friedo W. Dekker¹ | Merel van Diepen¹

¹Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands

²Department of Internal Medicine, Leiden University Medical Center, Leiden, The Netherlands

³Renal Research Institute, New York, USA

⁴Associazione Ipertensione Nefrologia Trapianto Renale (IPNET) Reggio Cal, Italy

⁵Department of Medical Informatics, ERA-EDTA Registry, Amsterdam UMC, University of Amsterdam, Amsterdam Public Health Institute, Amsterdam, The Netherlands

Correspondence

Ype de Jong, Department of Clinical Epidemiology, Leiden University Medical Center, PO Box 9600, Leiden, The Netherlands.

Email: y.de_jong@lumc.nl

Funding information

Dutch Kidney Foundation, Grant/Award Number: 16OKG12

Abstract

Over the past few years, a large number of prediction models have been published, often of poor methodological quality. Seemingly objective and straightforward, prediction models provide a risk estimate for the outcome of interest, usually based on readily available clinical information. Yet, using models of substandard methodological rigour, especially without external validation, may result in incorrect risk estimates and consequently misclassification. To assess and combat bias in prediction research the prediction model risk of bias assessment tool (PROBAST) was published in 2019. This risk of bias (ROB) tool includes four domains and 20 signalling questions highlighting methodological flaws, and provides guidance in assessing the applicability of the model. In this paper, the PROBAST will be discussed, along with an in-depth review of two commonly encountered pitfalls in prediction modelling that may induce bias: overfitting and composite endpoints. We illustrate the prevalence of potential bias in prediction models with a meta-review of 50 systematic reviews that used the PROBAST to appraise their included studies, thus including 1510 different studies on 2104 prediction models. All domains showed an unclear or high ROB; these results were markedly stable over time, highlighting the urgent need for attention on bias in prediction research. This article aims to do just that by providing (1) the clinician with tools to evaluate the (methodological) quality of a clinical prediction model, (2) the researcher working on a review with methods to appraise the included models, and (3) the researcher developing a model with suggestions to improve model quality.

KEYWORDS

clinical epidemiology, epidemiology, evidence-based medicine, medical education, meta-analysis

SUMMARY AT A GLANCE

Most published prediction models have limited clinical uptake, are not externally validated and come with methodological issues. The PROBAST (Prediction model Risk Of Bias ASsessment

Tool) guides the researcher writing a review, or the clinician interested in a model for risk calculation in a clinical setting. This review examines the aspects of bias in prediction research, and provides information on the prevalence of bias in published models.

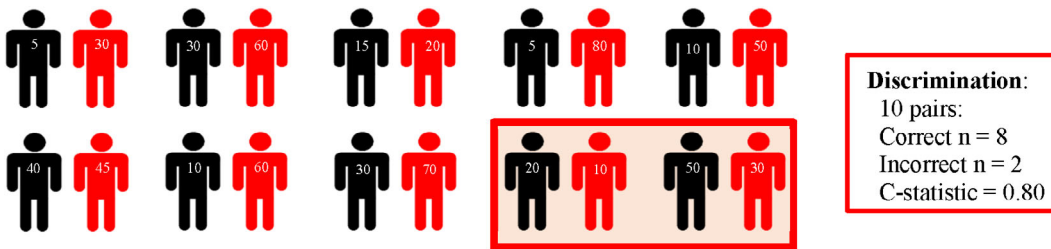
1 | INTRODUCTION AND BACKGROUND

Clinical prediction models are increasingly used for personalized medicine as these models inform on the diagnosis or the expected course of disease of individual patients. Two main groups of prediction models exist: models predicting the current presence or absence of a diagnosis (e.g., the WELLS score for screening for pulmonary embolism), and models predicting an outcome in the future (e.g., the KFRE-model for reaching end stage kidney disease in patients with chronic

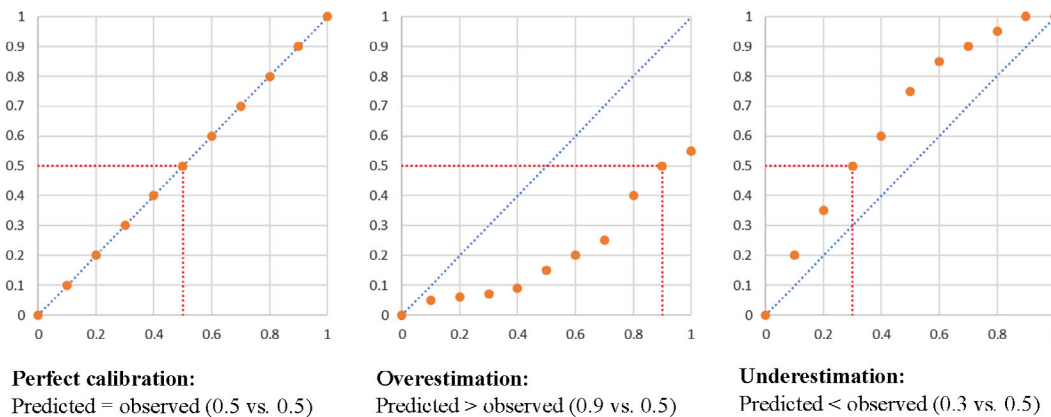
kidney disease). The main difference between these diagnostic- and prognostic models is the prediction timeframe, i.e. the time between the moment of prediction (i.e., baseline) and the occurrence of the outcome (respectively concurrent or in the future). After development, the predictive performance of models is typically assessed by discrimination and calibration: to what extent a model is able to differentiate between patients who reach the outcome and those who do not (discrimination), and to estimate a correct absolute risk (calibration) – concepts that are illustrated in more detail in Box 1.

BOX 1 Discrimination and calibration

Discrimination. Describes the models' ability to discriminate between events and nonevents (logistic models) or time-to-event (Cox proportional hazard). It is typically evaluated with the area under the receiver operating curve (AUC or AUROC for logistic models) or Harrel's C-statistic (for Cox proportional hazard models) for all possible pairs of nonevents and events. Below, we visualise the mechanism behind discrimination in a sample of 20 participants, consisting of 10 nonevents (black) and 10 events (red). The model assigned a higher probability to the events in 8 of the 10 pairs, but a lower probability in 2 of the 10 pairs (within the red box). Thus, the C-statistic is 0.80 (if this was a logistic model)



Calibration. Describes the relation between the observed risks within the population, and the predicted risks. Ideally, these risk would be equal in the entire range of predicted risks (from very low to very high risk patients). Typically, calibration is assessed by calibration-in-the-large, which is the average of the predicted and the average of the observed risks. Alternatively, calibration plots (below) can be constructed showing observed risk (y-axis) per decile of predicted risk (x-axis).



Perfect calibration:
Predicted = observed (0.5 vs. 0.5)

Overestimation:
Predicted > observed (0.9 vs. 0.5)

Underestimation:
Predicted < observed (0.3 vs. 0.5)

The increased interest and use of prediction models is reflected by the abundance of newly developed prediction models. For example, we recently identified 77 models developed for ischemic stroke,¹ and 42 models predicting kidney failure in patients with chronic kidney disease (CKD).² This already large number of models is exceeded by far in other fields such as cardiovascular disease (estimated at nearly 800 in 2015³) and pulmonology (models on chronic obstructive pulmonary disease estimated at more than 450 in 2019⁴). There are likely thousands of models in other fields published in bibliographic databases, and the number of models is increasing steadily (Figure 1). Unfortunately, most of these models have come with various methodological flaws, have limited clinical uptake and have not been externally validated, meaning that the performance was assessed in new patients.⁵ Although many models have found their way into everyday clinical practice, up until recently, no clear guidelines to assess a model's quality existed.

Bias is usually defined as the presence of a systematic error that may affect the study's validity. However, little empirical evidence on the effects of bias in prediction research exists, and it is unclear to what extent this definition of bias in the context of aetiological research is applicable to prediction. As the validity of a prediction model is tested in external validation, one way to look at bias is as a systematic difference between the model's estimated predictive performance in the development study, and the realised predictive performance in the validation study. In particular, methodological flaws in all stages of model development may result in a too optimistic estimate of predictive performance in the development cohort, which is not sustained in external settings.⁶ Optimism of predictive performance is not without risk, as flawed predicted risks will result in misclassification of the outcome (i.e., poor discrimination) and inaccurate risk estimation (i.e., poor calibration). Especially when the predicted risks of two separate outcomes are

compared, for example, the risk of ischemic stroke versus the risk of therapy related bleeding, misclassification is worrisome: over-prediction of ischemic stroke risk will result in an increased incidence of bleeding, and vice versa.⁷ Yet, the seeming simplicity and objectivity of prediction models – inserting clinical values that result in a risk of the outcome – is attractive to facilitate individualized patient care. Therefore, understanding of the potential pitfalls of prediction modelling is essential.

The aim of the present article is to provide readers with tools to appraise prediction models and assess their risk of bias (ROB) by discussing the recent publication of a ROB tool for prediction research. Next, using a meta-review approach, we will illustrate the prevalence of potential bias, and finally, we provide clear examples and illustrations of commonly encountered mistakes.

2 | APPRAISING PREDICTION RESEARCH: THE PROBAST

The Prediction model Risk Of Bias ASsessment Tool (PROBAST) was published in 2019⁸: it was designed as a general tool for critical appraisal of a single prediction model study, and for the use in systematic reviews of prediction models. An elaboration, discussing the different domains and signalling questions was published separately.⁶ It thus aimed to serve both the clinician that considers using a prediction model, and the researcher developing a model or including models in a systematic review or meta-analysis. The PROBAST contains two main domains: ROB and applicability. The ROB domain, which consists of four subdomains (participant selection, predictor selection, outcome definition and analysis), was defined by the authors as assessing what '(...) shortcomings in study design, conduct, or analysis could lead to systematically distorted estimates of a model's predictive performance'.⁶ The applicability domain addresses concerns regarding '(...) the applicability of a primary study to the review question can arise when the population, predictors, or outcomes of the study differ from those specified in the review question'.⁶ It consists of three subdomains (participant selection, predictor selection, outcome definition), and although this domain was developed for systematic reviews, the topics discussed can also apply to the use of prediction models in daily clinical practise. In total, the PROBAST contains 20 signalling questions which can be scored with low, unclear or high risk of bias, which in the end results in an overall judgement of low, unclear or high risk of bias and applicability – see Table 1. Below, we will discuss all subdomains of both ROB and applicability.

2.1 | Risk of bias: four subdomains

The first ROB subdomain, the *participants selection subdomain*, consists of two signalling questions and concerns the use of data sources and how participants were selected, with some study designs (e.g., observational cohorts or randomised controlled trials) at lower

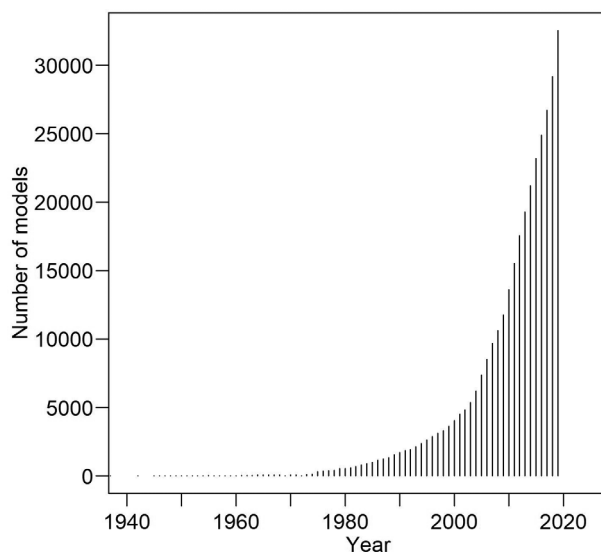


FIGURE 1 The increase in the number of prediction studies in PubMed (for the search string, see Data S1)

TABLE 1 The domains and signalling questions of the PROBAST for assessment of risk of bias and applicability. Data presented with permission of Wolff, coauthor of the PROBAST

Signalling questions				
	1. Participants	2. Predictors	3. Outcome	4. Analysis
Risk of bias	1.1. Were appropriate data sources used, for example, cohort, RCT, or nested case-control study data?	2.1. Were predictors defined and assessed in a similar way for all participants?	3.1. Was the outcome determined appropriately?	4.1. Were there a reasonable number of participants with the outcome?
	1.2. Were all inclusions and exclusions of participants appropriate?	2.2. Were predictor assessments made without knowledge of outcome data?	3.2. Was a prespecified or standard outcome definition used?	4.2. Were continuous and categorical predictors handled appropriately?
	-	2.3. Are all predictors available at the time the model is intended to be used?	3.3. Were predictors excluded from the outcome definition?	4.3. Were all enrolled participants included in the analysis?
	-	-	3.4. Was the outcome defined and determined in a similar way for all participants?	4.4. Were participants with missing data handled appropriately?
	-	-	3.5. Was the outcome determined without knowledge of predictor information?	4.5. Was selection of predictors based on univariable analysis avoided?
	-	-	3.6. Was the time interval between predictor assessment and outcome determination appropriate?	4.6. Were complexities in the data (e.g., censoring, competing risks, sampling of control participants) accounted for appropriately?
	-	-	-	4.7. Were relevant model performance measures evaluated appropriately?
	-	-	-	4.8. Were model overfitting, underfitting, and optimism in model performance accounted for?
	-	-	-	4.9. Do predictors and their assigned weights in the final model correspond to the results from the reported multivariable analysis?
	1. Participants	2. Predictors	3. Outcome	4. Analysis
Applicability	Included participants or setting does not match the review question	Definition, assessment, or timing of predictors does not match the review question	Its definition, timing, or determination does not match the review question	-

ROB, and some at higher risk (e.g., case-control studies). In case-control studies, cases are sampled and compared to a selection of controls; therefore the percentage of cases (and absolute risk of becoming a case) does not reflect the true absolute risk. In addition, the selection of participants should represent the target population: the model should be developed in a population that is similar to the population of its intended use. For example, models developed for late-stage chronic kidney disease (CKD) should be used with caution in early CKD, and vice-versa.⁹

The second ROB subdomain, the *predictors selection subdomain*, consists of three signalling questions and covers the sources of bias

that may arise due to the definition and measurement of the predictors. First, the predictors should be defined and assessed in the same way for all participants (e.g., an issue if the predictor ‘body weight’ was self-reported for some participants and measured for others). Furthermore, predictors should be assessed without knowledge of the outcome, that is, the outcome should be blinded when predictors are measured. In prospective research, predictor blinding is usually not an issue as the outcome is unknown at the moment predictors are established (e.g., future dialysis is not known when disease history is assessed). Finally, as a model should be usable in daily practise, all predictors should be available at the time of prediction. Although this

seems rather obvious, it is not uncommon to encounter models that include predictors available only after the moment of prediction, and thus not usable in a clinical setting.

The third ROB subdomain of the PROBAST is the *outcome subdomain*. It consists of six signalling questions that may point towards sources of bias in the outcome definition. First, the outcome should be determined appropriately: misclassification of the outcome will yield biased regression coefficients. Next, the outcome should be a prespecified or standard outcome, avoiding the risk of cherry-picking an outcome that yields the best model performance. Ideally, the outcome should not include predictors in any way. Though this sounds reasonable, this is often violated: for example, several kidney failure models define this outcome as an eGFR level below a certain threshold, but meanwhile included eGFR as predictor.² Incorporation bias, that is, inclusion of predictors in the outcome may result in overestimation of the relation between the predictor and the outcome, and thus an overly optimistic predictive performance.⁶ Clarity of outcome definitions is key: objective outcomes, such as a histological biopsy proving the presence or absence of disease, or survival versus nonsurvival, are less susceptible to bias than outcomes that require more interpretation of data. Obviously, if no objective outcome can be used, the outcome should be defined with such clarity that replication in a validation study, or application in the clinical field is possible. Next, this outcome should be defined and determined in a similar way for all participants. Consistent outcome definitions may be problematic in the setting of multi-centre studies, where centres use different methods to assess the outcome (e.g., the presence of an ischemic stroke using a CT-scan or MRI – whichever is available). If the endpoint is a composite of multiple outcomes, these individual components should be identical for all participants – a topic that will be discussed in more detail in Example 2. Lastly, the time interval between the predictor assessment and the outcome determination should be appropriate to capture the clinically relevant outcome.

The final ROB subdomain concerns the *analysis*. It consists of nine signalling questions which may point to flaws in the statistical methods.⁶ First, in Example 1, we will discuss the number of events in relation to overfitting, especially in a setting with a limited number of events and a large number of candidate predictors. Next, continuous predictors should not be dichotomized or categorised, as this will result in loss of information, which in turn may lead to risk estimates which are imprecise. Additionally, if categorisation cut-off points are based on the development dataset, for example, by using methods to identify the optimal cut-off point, the model will be overfit and biased. The next signalling question concerns the enrolment of participants in the analysis; excluding participants with outliers will likely result in bias. Furthermore, missing data – which is different from selective in- or exclusion – should be dealt with appropriately, preferably using multiple imputation instead of complete-case analysis.¹⁰ It should be noted that studies that do not mention missing data, or methods to deal with it, likely have conducted a complete-case analysis (and are thus at increased ROB) since most statistical packages automatically exclude participants with any missing information. Next, predictor selection and accounting for competing risks and censoring should be

done in a correct fashion, for example, when developing a prediction model for reaching end stage kidney disease in patients with chronic kidney disease, the competing risk of death is obvious and should be accounted for.² To evaluate the predictive performance – both in development and validation studies – performance measures such as discrimination and calibration should be presented in the study. Finally, all regression coefficients, including the baseline risk or model intercept, should be reported to allow the model to be used or externally validated. Many studies lack reporting baseline risk or model intercept. In addition, the presented regression coefficients should be in line with the coefficients of the final model, which may not be the case if authors retained only significant predictors of the multivariable analysis in their model, but did not re-estimate the coefficients in the smaller model. Alternatively, authors may present a risk score, where coefficients are rounded, thus losing information.

2.2 | Assessing applicability of prognostic models

In addition to ROB assessment, the PROBAST includes signalling questions to assess the applicability of existing prognostic models for systematic reviews. Though developed for the use of systematic reviews of prediction models, we believe these subdomains are also of relevance in daily clinical use or in the setting of an external validation study.

The first subdomain of the applicability section of the PROBAST, the *participant subdomain*, considers to which extent the population in the development studies matches the participants in the review – or the clinical setting. Development studies on populations from clinical trials, for example, need consideration: are the patients included in these trials comparable to the clinical setting, or are they healthier due to strict exclusion criteria? Furthermore, it is common for individuals enrolled in trials to be more involved in their health (i.e., healthy-user bias) which, in combination with the inclusion criteria, may result in limited external validity. Finally, models may be validated in a different specific population than intended: for example, we validated models for ischemic stroke in patients with CKD and dialysis patients.^{1,7} If the clinical rationale to do so is sound, and the model performs well, applicability may not be an issue, despite the difference in development and validation settings.

Differences in the heterogeneity of the study populations may result in differences in predictive performance. The next subdomain concerns the applicability of *predictors*, focussing on the differences in definitions, assessment or timing of predictors between the development study and the clinical setting. For example, a model using laboratory values as predictor is less applicable in a primary care setting, where blood sampling is not part of standard care. Or, alternatively, if the development study uses a predictor which value is assessed using specialised methods not routinely available, implementation in a clinical setting where this predictor is assessed using routine methods will likely result in poorer predictive performance.

The final subdomain concerns the applicability of the *outcome*. Again, as with the predictor definitions, differences in outcome

definition between validation studies or the outcome of interest in the clinical practise will likely influence the predictive performance. Concerns on applicability regarding the outcome may arise if the precision of the assessment methods of the outcome differs (similarly as with applicability of predictors), but composite outcomes may also result in applicability concerns (see Example 2).

EXAMPLE 1. In depth review of risk of bias due to predictor selection

The first step of model development concerns the selection of variables predicting an outcome – either predictors with a known aetiological relation or without (e.g., compare three different predictors for death: grey hair, advanced age, and telomere length – all three describe the relation of ageing with death, but with different degrees of causality). ROB regarding predictors is assessed in three subdomains of the PROBAST: the predictor definitions should be clear (predictor subdomain), predictors should not be used as or be part of the outcome (outcome subdomain), and the selection of predictors should be done in an appropriate manner (analysis). We will focus on the selection of candidate predictors for inclusion in the model, and discuss overfitting as a consequence of suboptimal selection methods in more detail.

It is common to encounter development studies that selected predictors from a list of candidate predictors by means of univariable selection: predictors with a statistically significant relation with the outcome are retained and included in the multivariable model. In univariable selection, predictors are selected based on their individual relation with the outcome, whilst in multivariable selection the strength of this relation is estimated in context with the other predictors. This univariable selection method may result in falsely rejecting predictors which may not have been statistically significant univariably but would be in a multivariable analysis, leading to poorer predictive performance. In addition, it is susceptible to singularities in the data leading to overfitting and thus optimism – which is a major problem in prediction research. Overfitting essentially describes a prediction model fitting the development data to precisely, as depicted in Figure 2. Although the model will show good predictive performance in this dataset, outside this sample, the performance will be poor, as the model does not reflect the underlying structure of the data. Overfitting can arise at many steps. We have discussed overfitting in relation to dichotomisation with data-driven cut-offs and data-driven selection methods, amongst others. Another commonly encountered cause is the sample size. The number of candidate predictors from which the final predictors are selected should be in proportion to the number of events or nonevents (whichever is smaller).

Several methods have been developed to reduce the risk of overfitting. For predictor selection, one method is to preselect predictors based on clinical knowledge or literature. Alternatively, data-driven

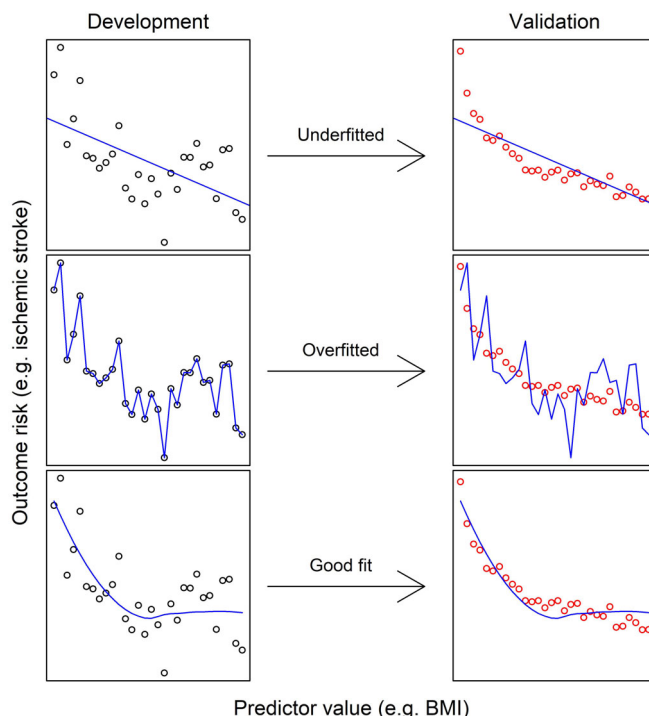


FIGURE 2 Model fitting illustrated. Different types of fit in candidate predictor selection, illustrated by two hypothetical samples of $n = 30$: a development cohort on the left, and a validation cohort on the right. Dots indicate the outcome risk for the predictor value (black dots in the development cohort; red dots in the validation cohort); the blue line indicates the fitted model. BMI; Body Mass Index

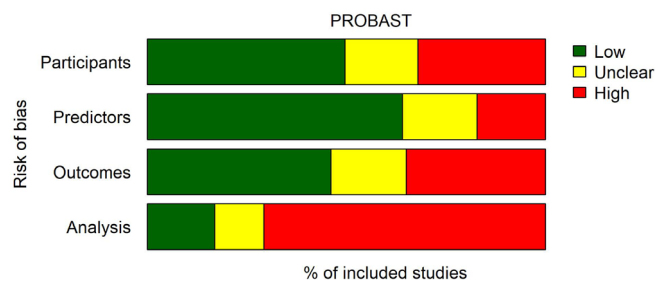


FIGURE 3 Aggregated overview of the risk of bias in 1039 prediction models with complete data (as assessed with the PROBAST in 50 systematic reviews)

methods such as backward selection may be performed. For this iterative method, a predefined cut-off for significance is used for in- and exclusion of predictors, and the regression components are re-estimated after each elimination. For adequate sample sizes, rules of thumb have been propagated: for Cox- and logistical modelling, 10 to 20 events per candidate predictor (e.g., meaning that, if the sample size exists of 40 events and 300 nonevents, only 2–4 candidate predictors can be used) have been suggested as appropriate sample sizes. Although the technical aspects of sample size calculations in prediction research are beyond the scope of this paper, this rule of thumb –

appealing because of its simplicity – has been under critique since it was proposed in the 1990s. New and more accurate methods for sample size calculation in prediction model development studies have been proposed recently.¹¹ Finally, shrinkage of the model's coefficients, or internal validation (i.e., statistically simulating an external validation in the development cohort) may be conducted. However, it can be argued that the most important step for assessing and recognising bias in prediction models is external validation: testing the capabilities of the model outside the population it was developed in.

EXAMPLE 2. Concerns of applicability and ROB due to composite outcomes

Composite outcomes are commonly used in prediction models: they allow developers to increase the number of events, and consequently the statistical power.¹ In the PROBAST, the authors acknowledge in two signalling questions that composite outcomes may lead to bias: as discussed above, a composite outcome should be defined beforehand and should not be adjusted based on the predictive performance. Next, the results of the individual components of the outcome should be combined in the same way for all participants. In addition, in our opinion, the clinical use should also be taken into consideration when developing a model. Take for example prediction

models on the risk of ischemic stroke – a risk that should be weighed against the risk of therapy related bleeding. Whilst inclusion of systemic embolus may be defensible from a clinical perspective, inclusion of haemorrhagic stroke in the composite outcome is odd, but encountered nonetheless.¹ Ideally, for these two entirely different outcomes, a predicted risk should be calculated and then weighed: if the risk of stroke is higher than bleeding, anti-coagulation may be prescribed or vice versa, withheld. Composite outcomes combining events that require different prevention strategies (e.g., death and dialysis in chronic kidney disease,² death and retransplantation after cardiac transplantation¹²) should be used with caution: if a high risk of the composite outcome is predicted, should the clinician counsel the patient for dialysis or retransplantation, or discuss conservative therapy?

2.3 | PROBAST: a meta-review on the prevalence of bias

To illustrate the ROB of currently available prediction models, we conducted a systematic literature search, identifying systematic reviews that used the PROBAST for risk of bias appraisal (methods detailed in Data S1). One year after its publication, after removal of duplicates, we

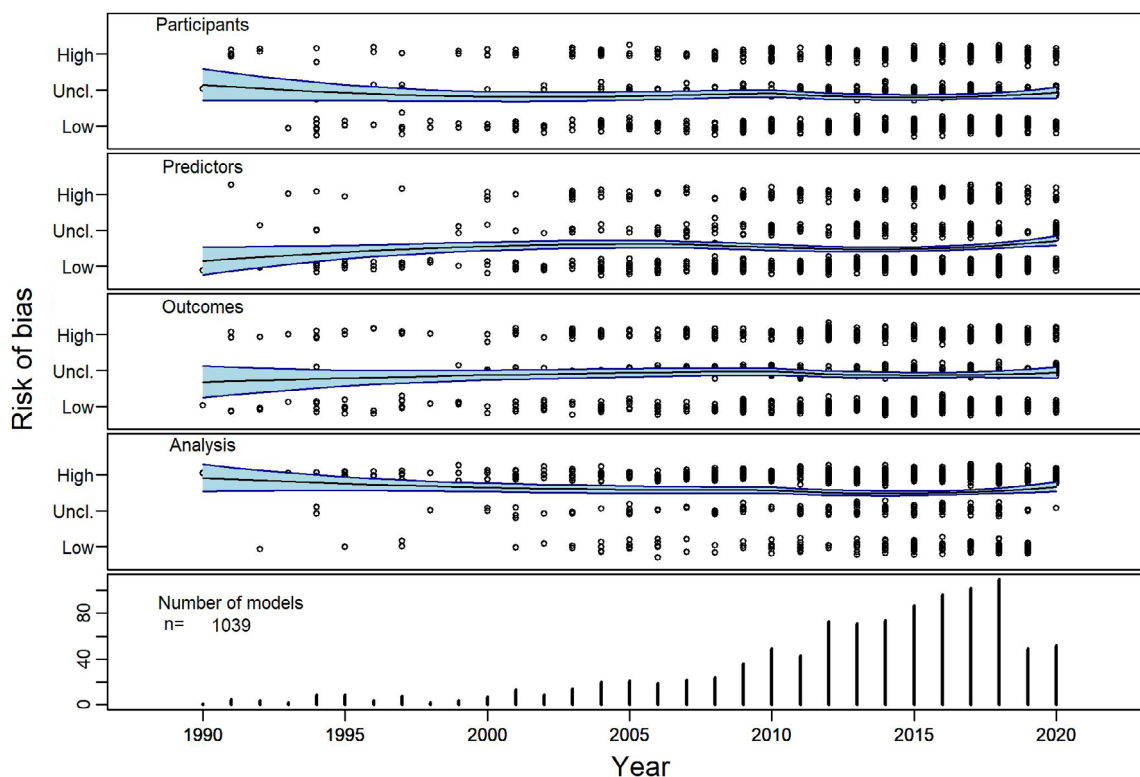


FIGURE 4 R Risk of bias of 1039 prediction models extracted from 50 systematic reviews with complete data as assessed with the PROBAST, stratified per year of publication and domain. Nine hundred and eighty five models presented information on the bias domains, and 560 presented information on the applicability domains. The trend is indicated by a fitted LOESS trendline with 95% confidence interval. For clarity, data points are jittered on the y-axis, by adding a Gaussian error with a standard deviation of 0.1

identified 151 articles that cited the PROBAST, of which 50 systematic reviews that used the PROBAST for ROB assessment, including in total 1510 studies on 2104 models (1458 development- and 646 validation studies), see Data S1, Table S1. All of the 50 reviews presented information on ROB; 17 did not present information on applicability. Eight reviews did not present data on bias per individual study and were therefore excluded for the analysis of ROB over time. In total, of the 2104 identified models in these 50 reviews, information of ROB per domain was presented for 1039 (47%) studies (see for these models Data S1, Table S2). Overall, ROB was judged by the authors of these reviews as high: of the 1039 studies with complete information on ROB, 25% scored a high ROB in participant selection, 18% scored a high risk on predictor selection, 31% of the studies scored a high risk on the domain outcome, whilst 69% scored a high ROB in the analysis domain (Figure 3, upper panel). When stratifying the ROB for the publication year of the included individual models (range 1966–2020), thus allowing visualisation of trends in ROB over time, two points become clear: (1) the recent increase in prediction models, with 72% (716) of the included models published in the last 10 years and (2) though the ROB for the participant and outcome domains decreased somewhat over time, the ROB in the analysis domain remained high (Figure 4).

2.4 | Perceived gaps

Although the PROBAST, and especially the accompanying elaboration article, covers most ROB and applicability issues that may be encountered, some topics receive relatively little attention, especially regarding applicability of models in a clinical- or research setting. Though most models present information on discrimination, information on calibration is often omitted: when adapting treatment based on the absolute risk estimate, the precision of this risk estimate is essential. We suggest that models offering incomplete information on calibration should be regarded at high concern for applicability in clinic. Another topic concerns the prediction horizon: the duration of time in which the outcome could occur. The length of this prediction horizon is obviously dependent on the outcome: early warning scores predicting adverse outcomes during hospital stays will have shorter prediction horizons than risk of death due to diabetes. Regardless, the prediction horizon should be predefined (e.g., respectively at 3 days or 5 years), else clinical use or external validation is limited, as it is uncertain to what timeframe the predicted risks apply. Finally, models should present their risk estimation as an absolute risk (i.e., cumulative incidence), ideally corrected for competing risks. It is not uncommon however to encounter models presenting a hazard rate or event rate (i.e., events per person-years) instead of an absolute risk, making risk estimation and calibration cumbersome if not impossible.

3 | SUMMARY AND CONCLUSIONS

Prediction models are promising for individualized medicine but the overwhelming quantity and often poor quality have limited

implementation in clinical practise. The PROBAST, a checklist designed to estimate ROB and assess applicability, helps the reader to determine the quality of a prediction model. This tool was well-received, as demonstrated with the already large number of systematic reviews using it just 1 year after publication. By analysing these systematic reviews, we were able to illustrate the abundance of prediction models, and demonstrate the trends in ROB and application over time – especially so in the analysis domain. Our review of the PROBAST, together with the elaboration paper by the authors, may serve both the clinician looking to implement a model in daily practise, the researcher that aims to develop or validate a model, and the researcher conducting a systematic review on prediction models.

ACKNOWLEDGEMENT

This study was supported by a grant from the Dutch Kidney Foundation (16OKG12).

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

ORCID

Ype de Jong  <https://orcid.org/0000-0002-7805-1508>

REFERENCES

- de Jong Y, Ramspek CL, van der Endt VHW, et al. A systematic review and external validation of stroke prediction models demonstrates poor performance in dialysis patients. *J Clin Epidemiol* 2020;123:69–79. doi: <https://doi.org/10.1016/j.jclinepi.2020.03.015>
- Ramspek CL, de Jong Y, Dekker FW, et al. Towards the best kidney failure prediction tool: a systematic review and selection aid. *Nephrol, Dial, Transplant*. 2020;35(9):1527–1538. <https://doi.org/10.1093/ndt/gfz018>
- Wessler BS, Lai Yh L, Kramer W, 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):368–375. <https://doi.org/10.1161/circoutcomes.115.001693>
- Bellou V, Belbasis L, Konstantinidis AK, et al. Prognostic models for outcome prediction in patients with chronic obstructive pulmonary disease: systematic review and critical appraisal. *BMJ-Br Med J*. 2019; 367:15. <https://doi.org/10.1136/bmj.l5358>
- Ramspek CL, Jager KJ, Dekker FW, et al. External validation of prognostic models: what, why, how, when and where? *Clin Kidney J*. 2021;14(1):49–58. <https://doi.org/10.1093/ckj/sfaa188>
- Moons KGM, Wolff RF, Riley RD, 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. <https://doi.org/10.7326/m18-1377>
- de Jong Y, Fu EL, van Diepen M, et al. Validation of risk scores for ischaemic stroke in atrial fibrillation across the spectrum of kidney function. *Eur Heart J*. 2021;42(15):1476–1485. <https://doi.org/10.1093/eurheartj/ehab059>
- Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;170(1):51–58. <https://doi.org/10.7326/m18-1376>

9. Ramspek CL, Evans M, Wanner C, et al. Kidney failure prediction models: a comprehensive external validation study in patients with advanced CKD. *J Am Soc Nephrol*. 2021;32(5):1174–1186. <https://doi.org/10.1681/asn.2020071077>
10. de Goeij MC, van Diepen M, Jager KJ, et al. Multiple imputation: dealing with missing data. *Nephrol, Dial, Transplant*. 2013;28(10):2415–2420. <https://doi.org/10.1093/ndt/gft221>
11. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ-Br Med J*. 2020;368:m441. <https://doi.org/10.1136/bmj.m441>
12. Aleksova N, Alba AC, Molinero VM, et al. Risk prediction models for survival after heart transplantation: a systematic review. *Am J Transplant*. 2020;20(4):1137–1151. <https://doi.org/10.1111/ajt.15708>

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: de Jong Y, Ramspek CL, Zoccali C, Jager KJ, Dekker FW, van Diepen M. Appraising prediction research: a guide and meta-review on bias and applicability assessment using the Prediction model Risk Of Bias ASsessment Tool (PROBAST). *Nephrology*. 2021;26(12): 939–947. <https://doi.org/10.1111/nep.13913>