Annals of Internal Medicine RESEARCH AND REPORTING METHODS

PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies

Robert F. Wolff, MD*; Karel G.M. Moons, PhD*; Richard D. Riley, PhD; Penny F. Whiting, PhD; Marie Westwood, PhD; Gary S. Collins, PhD; Johannes B. Reitsma, MD, PhD; Jos Kleijnen, MD, PhD; and Sue Mallett, DPhil; for the PROBAST Group†

Clinical prediction models combine multiple predictors to estimate risk for the presence of a particular condition (diagnostic models) or the occurrence of a certain event in the future (prognostic models).

PROBAST (Prediction model Risk Of Bias ASsessment Tool), a tool for assessing the risk of bias (ROB) and applicability of diagnostic and prognostic prediction model studies, was developed by a steering group that considered existing ROB tools and reporting guidelines. The tool was informed by a Delphi procedure involving 38 experts and was refined through piloting.

PROBAST is organized into the following 4 domains: participants, predictors, outcome, and analysis. These domains contain a total of 20 signaling questions to facilitate structured judgment of ROB, which was defined to occur when shortcomings in study design, conduct, or analysis lead to systematically distorted estimates of model predictive performance. PROBAST enables a focused and transparent approach to assessing the ROB and applicability of studies that develop, validate, or update prediction models for individualized predictions.

Although PROBAST was designed for systematic reviews, it can be used more generally in critical appraisal of prediction model studies. Potential users include organizations supporting decision making, researchers and clinicians who are interested in evidence-based medicine or involved in guideline development, journal editors, and manuscript reviewers.

Ann Intern Med. 2019;170:51-58. doi:10.7326/M18-1376 **Annals.org**

For author affiliations, see end of text.

* Drs. Wolff and Moons contributed equally to this work.

† For members of the PROBAST Group, see the **Appendix** (available at Annals.org).

Prediction relates to estimating the probability of something currently unknown. In the context of medical research, prediction typically concerns either diagnosis (probability of a certain condition being present but not yet detected) or prognosis (probability of an outcome developing in the future) (1–3). Prognosis applies not only to sick persons or those with an established diagnosis but also to, for example, pregnant women at risk for diabetes (4). Prediction research includes predictor finding studies, prediction model studies (development, validation, and extending or updating), and prediction model impact studies (1).

Predictor finding studies (also known as risk factor or prognostic factor studies) aim to identify which predictors (such as age, disease stage, or biomarkers) independently contribute to the prediction of a diagnostic or prognostic outcome (1, 5).

Prediction model studies aim to develop, validate, or update (for example, extend) a multivariable prediction model. A prediction model uses multiple predictors in combination to estimate probabilities to inform and often guide individual care (2, 6, 7). These models can predict an individual's probability of either currently having a particular outcome or disease (diagnostic prediction model) or having a particular outcome in the future (prognostic prediction model). Both types of model are widely used in various medical domains and settings (8–10), as evidenced by the large number of models developed in cancer (11, 12), neurology (13, 14), and cardiovascular disease (15). Prediction models are sometimes described as risk prediction models, predictive models, prediction indices or rules, or risk scores (2, 7). An example is QRISK2 for predicting cardiovascular risk (16).

Prediction model impact studies evaluate the effect of using a model to guide patient care compared with not using such a model. They use a comparative design, such as a randomized trial, to study the model's effect on clinical decision making, patient outcomes, or costs of care (1).

Systematic reviews have a key role in evidence-based medicine and the development of clinical guidelines (17–19). They are considered to provide the most reliable form of evidence for the effects of an intervention or diagnostic test (20, 21). Systematic reviews of prediction models are a relatively new and evolving area but are increasingly undertaken to systematically identify, appraise, and summarize evidence on the performance of prediction models (1, 6, 22).

Assessing the quality of included studies is a crucial step in any systematic review (20, 21). The QUIPS (Quality In Prognosis Studies) tool has been developed to assess risk of bias (ROB) in predictor finding (prognostic factor) studies (23). Researchers can use the revised Cochrane ROB Tool (ROB 2.0) (24) to investigate the methodological quality of prediction model impact studies that use a randomized comparative design, or ROBINS-I (Risk Of Bias In Nonrandomized Studies of Interventions) for those that use a nonrandomized comparative design (25). As more prediction model studies and systematic reviews of such studies are used as evidence for clinical guidance, a tool facilitating quality assessment for individual prediction model studies is urgently needed.

See also:

Web-Only
Explanation and Elaboration
Supplement

Box 1. Types of diagnostic and prognostic modeling studies or reports addressed by PROBAST.

Prediction model development without external validation: **These studies aim to develop prognostic or diagnostic prediction models from a specific development data set. They aim to identify the important predictors of the outcome under study, assign weights (e.g., regression coefficients) to each predictor using some form of multivariable analysis, develop a prediction model to be used for individualized predictions, and quantify the predictive performance of that model in the development set. Sometimes, model development studies may also focus on adding new predictors to established predictors. In any prediction model study, overfitting may occur, particularly in small data sets. Hence, development studies should include some form of resampling or "internal validation" (internal because the same data are used for both development and internal validation), such as bootstrapping or cross-validation. These methods quantify any optimism (bias) in the predictive performance of the developed model.**

Prediction model development with external validation: **These studies have the same aim as the previous type, but the development of the model is followed by quantifying its predictive performance in data external to the development sample (i.e., from different participants). These data may be collected by the same investigators, commonly using the same predictor and outcome definitions and measurements but sampled from a later time period (temporal validation); by other investigators in another hospital or country, sometimes using different definitions and measurements (geographic validation); in similar participants but from an intentionally chosen different setting (e.g., a model developed in secondary care and tested in similar participants from primary care); or even in other types of participants (e.g., a model developed in adults and tested in children). Randomly splitting a single data set into a development and a validation data set is often erroneously referred to as a form of external validation but actually is an inefficient form of internal validation, because the 2 data sets created in this way differ only by chance and the sample size of model development is reduced. When a model predicts poorly when validated in other data, a model validation can be followed by adjusting (or updating the existing model [e.g., by recalibration of the baseline risk or hazard or adjusting the weights of the predictors in the model]) to the validation data set at hand and even by extending the model by adding new predictors to the existing model. In both situations, a new model is in fact being developed after the external validation of the existing model.**

Prediction model external validation: **These studies aim to assess the predictive performance of existing prediction models using data** ***external*** **to the development sample (i.e., from different participants).**

Adopted from the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) and CHARMS (CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies) guidance (7, 26). PROBAST = Prediction model Risk Of Bias ASsessment Tool.

We present PROBAST (Prediction model Risk Of Bias ASsessment Tool), a tool to assess the ROB and concerns regarding the applicability of diagnostic and prognostic prediction model studies. PROBAST can be used to assess studies of model development and model validation, including those updating a prediction model (Box 1 [26]). We refer to the accompanying explanation and elaboration document (27), available at Annals.org, for detailed explanations of how to use PROBAST and how to judge ROB and applicability.

METHODS: DEVELOPMENT OF PROBAST

Development of PROBAST was based on a 4-stage approach for developing health research reporting guidelines: define the scope, review the evidence base, use a Web-based Delphi procedure, and refine the tool through piloting (28). Guidelines explicitly aimed at the development of quality assessment tools were not available at the time (29).

Development Stage 1: Scope and Definitions

A steering group of 9 experts in prediction model studies and development of quality assessment tools agreed on key features of the desired scope of PROBAST. A panel of 38 experts with different backgrounds further refined the scope during the Web-based Delphi procedure.

PROBAST was designed mainly to assess primary studies included in a systematic review. The group agreed that PROBAST would assess both *risk of bias* and *concerns regarding applicability* of a study evaluating a multivariable prediction model to be used for individualized diagnosis or prognosis. A domain-based structure was adopted, similar to that used in other ROB tools, such as ROB 2.0 (24), ROBINS-I (25), QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies 2) (30), and ROBIS (31).

We agreed that PROBAST should cover primary studies that develop, validate, or update multivariable prediction models aiming to make individualized predictions of a diagnostic or prognostic outcome (Box 1). Studies that use multivariable modeling techniques to identify predictors (such as risk or prognostic factors) associated with an outcome but do not attempt to develop, validate, or update a model for making individualized predictions are not covered by PROBAST (5). Therefore, PROBAST is not intended for predictor finding studies or prediction model impact studies.

Studies of diagnostic and prognostic models often use different terms for predictors and outcomes (Box 2). A multivariable prediction model is defined as any combination or equation of 2 or more predictors for estimating probability or risk for an individual (6, 7, 32–34).

Development Stage 2: Review of Evidence

We used the following 3 approaches to build an evidence base to inform the development of PROBAST: identifying relevant methodological reviews in the area of prediction model research (November 2012 to January 2013), asking members of the steering group to identify relevant methodological studies (January 2013 to March 2013), and using the Delphi procedure to ask members of the wider group to identify additional evidence (February 2012 to July 2014).

Identified literature was used to guide the scope and produce an initial list of signaling questions to consider for inclusion in PROBAST (1, 2, 5–7, 26, 33–40). We grouped signaling questions into common themes to identify possible domains. Additional literature provided as part of the Web-based surveys informed development of the explanation and elaboration document.

Development Stage 3: Web-Based Delphi Procedure

We used a modified Delphi process to gain feedback and agreement on the scope, structure, and content of PROBAST. Web-based surveys were developed to gather structured feedback for each round. The 38-member Delphi group comprised methodological experts in prediction model research and development of quality assessment tools, experienced systematic reviewers, commissioners, and representatives of reimbursement agencies. We included various stakeholders to ensure that the views of end users, methodological experts, and decision makers were represented.

The Delphi process consisted of 7 rounds. Round 1 asked about the scope of the tool, and participants agreed to focus on prediction model studies and follow a domain-based structure. Round 2 aimed to identify relevant domains and agree on which to include. The signaling questions for domains were refined in rounds 3 to 5. Respondents used a 1-to-5 Likert scale to rate each proposed signaling question for inclusion. They could also suggest rephrasing, provide supporting evidence (such as references to relevant studies), and suggest missing signaling questions. Round 6 refined the domains and introduced further optional guidance for using PROBAST. In the last round, participants received the agreed draft version of PROBAST and had the opportunity to provide any final feedback.

Development Stage 4: Piloting and Refining the Tool

We held 6 workshops on PROBAST at consecutive annual Cochrane Colloquia (Quebec, Canada, 2013; Hyderabad, India, 2014; Vienna, Austria, 2015; Seoul, South Korea, 2016; Cape Town, South Africa, 2017; and Edinburgh, United Kingdom, 2018). We also held numerous consecutive workshops with MSc and PhD students (for example, the master's program in epidemiology at Utrecht University [Utrecht, the Netherlands] and the Evidence-Based Health Care program at Oxford University [Oxford, United Kingdom]). In these workshops, we piloted the then-current version of PROBAST to gather feedback on practical issues asso-

Box 2. Differences between diagnostic and prognostic prediction model studies.

Diagnostic **prediction models aim to estimate the probability that a target condition measured using a reference standard (referred to as an "outcome" in PROBAST) is currently present or absent within an individual. In diagnostic prediction model studies, the prediction is for an outcome already present, so the preferred design is a cross-sectional study. However, sometimes follow-up is used as part of the reference test to determine whether the target condition (e.g., a disease) is present at the moment of prediction.**

Prognostic **prediction models estimate whether an individual will experience a specific event or outcome in the future within a certain time period, ranging from minutes to hours, days, weeks, months, or years; the relationship is always longitudinal.**

Despite the different timing of the predicted outcome, diagnostic and prognostic prediction models have many similarities, including the following:

- **The type of outcome is often binary (whether the target condition is present or not present, or an outcome event will or will not occur in the future).**
- **The key interest is to estimate the probability of an outcome being present or occurring in the future based on multiple predictors with the purpose of informing individuals and guiding decision making.**
- **The same challenges occur when developing or validating multivariable prediction models. The same measures for assessing the predictive performance of the model can be used, although diagnostic models more frequently extend assessment of predictive performance to focus on thresholds of clinical relevance.**

There are also various differences in terminology between diagnostic and prognostic model studies, including the following:

Diagnostic Prediction Model Study	Prognostic Prediction Model Study
Predictors	
Diagnostic tests or index tests	Prognostic factors or prognostic indicators
Outcome	
Reference standard used to assess or verify the presence/absence of the target condition	Event (whether an event will occur in the future); event measurement
Missing outcome assessment	
Partial verification, lost to follow-up	Lost to follow-up and censoring

PROBAST = Prediction model Risk Of Bias ASsessment Tool.

Table 1. Four Steps in PROBAST

Step	Task	When to Complete
1	Specify your systematic review question(s)	Once per systematic review
2	Classify the type of prediction model evaluation	Once for each model of interest in each publication being assessed, for each relevant outcome
3	Assess risk of bias and applicability (per domain)	Once for each development and validation of each distinct prediction model in a publication
4	Overall judgment of risk of bias and applicability	Once for each development and validation of each distinct prediction model in a publication

PROBAST = Prediction model Risk Of Bias ASsessment Tool.

ciated with using the tool so that we could further refine and subsequently validate it. Finally, more than 50 review groups have already piloted PROBAST versions, including the final version, in their reviews. Topics included cancer, cardiology, endocrinology, pulmonology, and orthopedics.

All feedback received from these initiatives was used to further inform the content and structure of PROBAST, wording of the signaling questions, and content of the guidance documents (27).

RESULTS: THE PROBAST TOOL

What Does PROBAST Assess?

PROBAST assesses both *risk of bias* and *concerns regarding applicability* of primary studies that developed or validated multivariable prediction models for diagnosis or prognosis (Boxes 1 and 2).

Development of a prediction model can include adding new predictors to an existing prediction model. Similarly, validation of an existing model can be accompanied by updating and extending the model—that is, development of a new model. PROBAST applies to both situations (Box 1).

Target Users

Although PROBAST was designed for use in systematic reviews, it can be used more generally in critical appraisal of prediction model studies. Potential users of PROBAST include organizations supporting decision making (such as the National Institute for Health and Care Excellence and the Institute for Quality and Efficiency in Health Care); researchers and clinicians who are interested in evidence-based medicine or involved in guideline development; and journal editors, manuscript reviewers, and readers who want to critically appraise prediction model studies.

Definition of ROB and Applicability

Bias is usually defined as the presence of systematic error in a study that leads to distorted or flawed results and hampers the study's internal validity. In prediction model development and validation, known features exist that make a study at ROB, although *empirical* evidence showing the most important sources of bias is limited. We define ROB to occur when shortcomings in the study design, conduct, or analysis lead to systematically distorted estimates of model predictive performance. Model predictive performance is typically evaluated using measures of calibration and discrimination, and sometimes (notably in diagnostic model studies) classification (7). Thinking about how a hypothetical prediction model study that is methodologically robust would have been designed, conducted, and analyzed helps to understand bias in study estimates of model predictive performance. Many sources of bias identified in other medical research areas are also relevant to prediction model studies, such as blinding of outcome assessors to other study features and use of consistent definitions and measurements for predictors and outcomes within the study.

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. Such concerns may arise when participants in the prediction model study are from a different medical setting from the population defined in the review question—for example, a study that enrolled patients from a hospital setting while the review question specifically relates to patients in primary care. The reported prediction model discrimination and calibration may not be applicable because patients in hospital settings typically have more severe disease than those in primary care (41, 42).

When eligibility criteria, predictors, and outcomes of the primary studies directly match a systematic review question, no concerns regarding applicability will arise. However, the inclusion criteria of a systematic review are typically broader than the focus of the review question. Broader inclusion criteria allow for variation in the searching of the primary studies and thus require careful assessment of each primary study's applicability to the actual review question (7, 27).

Types of Prediction Model Study

A primary study identified as relevant for the review may include the development, validation, or update of 1 or more prediction models. For each study, a PROBAST assessment should be completed for each distinct model that is developed, validated, or updated for making individualized predictions relevant to the systematic review question.

PROBAST includes 4 steps (Table 1). The tool is in the Supplement (available at Annals.org). We stress the importance of the accompanying paper (27), which provides detailed explanations and guidance for completing each step.

Step 1: Specify Your Systematic Review Question

Assessors are first asked to report their systematic review question in terms of intended use of the model, targeted participants, predictors used in the modeling, and predicted outcome. Existing guidance (CHARMS [CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies]) can help reviewers define a clear and focused review question (22, 26).

Step 2: Classify the Type of Prediction Model Evaluation

Different signaling questions apply to different types of prediction model evaluation. For each model assessment, reviewers classify a model as "development only," "development and validation in the same publication," or "validation only." When a publication focuses on creating a model by adding 1 or more new predictors to established predictors (or an established model), "development only" should be used. When a publication focuses on validating an existing model in other data and then updating (adjusting or extending) the model such that a new model is actually being developed, "development and validation in the same publication" should be used. Note again that a single publication may address more than 1 model of interest.

Step 3: Assess ROB and Applicability

Step 3 aims to identify areas where bias may be introduced into the prediction model study or where concerns regarding applicability may exist. It involves assessment of the following 4 domains to cover key aspects of prediction model studies: participants, predictors, outcome, and analysis. The ROB component of each domain comprises 4 sections: information used to support the judgment, 2 to 9 signaling questions (20 total across domains), judgment of ROB, and rationale for the judgment (Table 2).

In the support for judgment box, assessors can record the information used to answer the signaling questions. Signaling questions are answered as "yes," "probably yes," "probably no," "no," or "no information." Risk of bias is judged as low, high, or unclear. All signaling questions are phrased so that "yes" indicates absence of bias. Any signaling question answered as "no" or "probably no" flags the potential for bias; assessors will need to use their own judgment to determine whether the domain should be rated as high, low, or unclear ROB. A "no" answer does not automatically result in a high ROB rating. The "no information" category should be used only when reported information is in-

Table 2. PROBAST: Summary of Step 3—Assessment of Risk of Bias and Concerns Regarding Applicability*

1. Participants	2. Predictors	3. Outcome	4. Analysis
Signaling questions			
1.1. Were appropriate data sources used, e.g., 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?†
ROB			
Selection of participants	Predictors or their assessment	Outcome or its determination	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	–

RCT = randomized controlled trial; ROB = risk of bias.
* For further details, please see the explanation and elaboration document (27), available at Annals.org, and www.probast.org. Signaling questions are answered as yes, probably yes, probably no, no, or no information. ROB and concerns for applicability are rated as low, high, or unclear.
† Development studies only.

Table 3. Suggested Tabular Presentation for PROBAST Results*

Study	ROB				Applicability			Overall	
	Participants	Predictors	Outcome	Analysis	Participants	Predictors	Outcome	ROB	Applicability
1	+	−	?	+	+	+	+	−	+
2	+	+	+	+	+	+	+	+	+
3	+	+	+	?	−	+	+	?	−
4	−	?	?	−	+	+	−	−	−
5	+	+	+	+	+	?	+	+	?
6	+	+	+	+	?	+	?	+	?
7	?	?	+	?	+	+	+	?	+
8	+	+	+	+	+	+	+	+	+

PROBAST = Prediction model Risk Of Bias ASsessment Tool; ROB = risk of bias.
* + indicates low ROB/low concern regarding applicability; − indicates high ROB/high concern regarding applicability; and ? indicates unclear ROB/unclear concern regarding applicability.

sufficient to permit a judgment. When the rationale is recorded, the ROB rating will be transparent and, where necessary, will facilitate discussion among review authors completing assessments independently.

The first 3 domains are also rated for concern regarding applicability (low, high, or unclear) to the review question defined in step 1. Concerns regarding applicability are rated similarly to ROB, but without signaling questions.

All domains should be completed separately for each evaluation of a distinct model in each study. A team completing a PROBAST assessment likely needs both subject and methodological expertise. The explanation and elaboration document (27) and www.probast.org provide further details on how to score ROB and applicability concerns. *Domain 1 (Participants)* covers potential sources of bias and applicability concerns related to participant selection methods and data sources (for example, study designs); 2 signaling questions support ROB assessment. *Domain 2 (Predictors)* covers potential sources of bias and applicability concerns related to the definition and measurement of predictors evaluated for inclusion in the model; 3 signaling questions support ROB assessment. *Domain 3 (Outcome)* covers potential sources of bias and applicability concerns related to the definition and measurement of the outcome predicted by the model; 6 signaling questions support ROB assessment. *Domain 4 (Analysis)* covers potential sources of bias in the statistical analysis methods. It assesses aspects related to the choice of analysis method and whether key statistical considerations (for example, missing data) were correctly addressed, and 9 signaling questions support ROB assessment.

Table 2 presents an overview of step 3. Detailed examples of how to answer signaling questions and judge domains can be found in the explanation and elaboration document (27) and on www.probast.org.

Step 4: Overall Judgment

On the basis of the ROB classifications for each domain in step 3, assessors should judge the *overall* ROB of the prediction model as low, high, or unclear. We recommend rating the prediction model as having low ROB if no relevant shortcomings were identified in the ROB assessment—that is, all domains had low ROB. If at least 1 domain had high ROB, an overall judgment of high ROB should be used. Similarly, unclear ROB should be assigned if unclear ROB was noted in at least 1 domain and all other domains had low ROB.

However, if a prediction model was developed without any external validation on different participants, downgrading to high ROB should still be considered even if all 4 domains had low ROB, unless the model development was based on a very large data set or included some form of internal validation. The explanation and elaboration document (27) provides further details.

Based on the applicability classifications for each domain in step 3, an overall judgment about concerns regarding applicability of the prediction model is needed. A decision of "low concern" should be reached only if all domains showed low concern regarding applicability. Similarly, if 1 or more domains were judged to have high concern, the overall judgment should be "high concern." "Unclear concern regarding applicability" should be reached only if 1 or more domains were judged as "unclear" in applicability and all other domains were rated to have "low concern."

The accompanying explanation and elaboration document (27) and www.probast.org give detailed explanation and examples of how to judge the overall ROB and concerns regarding applicability. Table 3 suggests a way to present the results of the PROBAST assessments.

DISCUSSION

Assessment of the quality of included studies is an essential component of all systematic reviews and evidence syntheses. Systematic reviews of prediction model studies are a rapidly evolving area (22). As more prediction model studies and systematic reviews of such studies enter the evidence base, a tool facilitating quality assessment for individual prediction model studies is urgently needed. To our knowledge, PROBAST is the first rigorously developed tool designed specifically to assess the quality of prediction model studies for development, validation, or updating of both diagnostic and prognostic models, regardless of the medical domain, type of outcome, predictors, or statistical technique used.

We adopted a domain-based structure similar to that used in other recently developed tools, such as ROB 2.0 (24), QUADAS-2 for diagnostic accuracy studies (30), ROBINS-I for nonrandomized studies (25), and ROBIS for systematic reviews (31). All stages of PROBAST development included a wide range of stakeholders, and we started piloting the tool in early versions to allow incorporation of feedback from direct reviewer experience into the final tool. We feel that these 2 features have resulted in a tool that is both methodologically sound and user-friendly.

Potential users of PROBAST include systematic review authors, health care decision makers, and researchers and clinicians who are interested in evidence-based medicine or involved in guideline development, as well as journal editors and manuscript reviewers.

The explanation and elaboration document (27) provides explicit guidance and an explanation of how to use PROBAST. Researchers seeking to understand and use PROBAST should always read the accompanying document in conjunction with the current article. A multidisciplinary team with both subject and methodological expertise should assess prediction model studies.

As with other ROB and reporting guidelines in medical research, PROBAST and its guidance will require updating as methods for prediction model studies develop. We recommend downloading the latest version of PROBAST and accompanying guidance, including detailed examples, from the Web site (www.probast.org).

From Kleijnen Systematic Reviews, York, United Kingdom (R.F.W., M.W.); Julius Center for Health Sciences and Primary Care and Cochrane Netherlands, University Medical Center Utrecht, Utrecht University, Utrecht, the Netherlands (K.G.M., J.B.R.); Centre for Prognosis Research, Research Institute for Primary Care and Health Sciences, Keele University, Keele, United Kingdom (R.D.R.); Medical School of the University of Bristol and National Institute for Health Research Collaboration for Leadership in Applied Health Research and Care West, University Hospitals Bristol National Health Service Foundation Trust, Bristol, United Kingdom (P.F.W.); Centre for Statistics in Medicine, University of Oxford, Oxford, United Kingdom (G.S.C.); Kleijnen Systematic Reviews, York, United Kingdom, and School for Public Health and Primary Care, Maastricht University, Maastricht, the Netherlands (J.K.); and Institute of Applied Health Research, National Institute for Health Research Birmingham Biomedical Research Centre, College of Medical and Dental Sciences, University of Birmingham, Birmingham, United Kingdom (S.M.).

Disclaimer: This report presents independent research supported by the National Institute for Health Research (NIHR). The views and opinions expressed in this publication are those of the authors and do not necessarily reflect those of the National Health Service (NHS), the NIHR, or the Department of Health and Social Care.

Acknowledgment: The authors thank the members of the Delphi panel (**Appendix**, available at Annals.org) for their valuable input and all testers, especially Cordula Braun, Johanna A.A.G. Damen, Paul Glasziou, Pauline Heus, Lotty Hooft, and Romin Pajouheshnia, for providing feedback on PROBAST. They also thank Janine Ross and Steven Duffy for support in managing the references.

Financial Support: Drs. Moons and Reitsma received financial support from the Netherlands Organisation for Scientific Research (ZONMW 918.10.615 and 91208004). Dr. Riley is a member of the Evidence Synthesis Working Group funded by the NIHR School for Primary Care Research (project 390). Dr. Whiting (time) was supported by the NIHR Collaboration for Leadership in Applied Health Research and Care West at University Hospitals Bristol NHS Foundation Trust. Dr. Collins was supported by the NIHR Biomedical Research Centre, Oxford. Dr. Mallett is supported by NIHR Birmingham Biomedical Research Centre at the University Hospitals Birmingham NHS Foundation Trust and the University of Birmingham. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Disclosures: Dr. Wolff reports grants from Bayer, Biogen, Pfizer, UCB, Amgen, BioMarin, Grünenthal, Chiesi, and TESARO outside the submitted work. Dr. Westwood reports grants from Bayer, Biogen, Pfizer, UCB, Amgen, BioMarin, Grünenthal, Chiesi, and TESARO outside the submitted work. Dr. Kleijnen reports grants from Bayer, Biogen, Pfizer, UCB, Amgen, BioMarin, Grünenthal, Chiesi, and TESARO outside the submitted work. Authors not named here have disclosed no conflicts of interest. Disclosures can also be viewed at www.acponline.org/authors/icmje/ConflictOfInterestForms.do?msNum=M18-1376.

Corresponding Author: Robert F. Wolff, MD, Kleijnen Systematic Reviews Ltd, Unit 6, Escrick Business Park, Riccall Road, Escrick, York YO19 6FD, United Kingdom; e-mail, robert@systematic-reviews.com.

Current author addresses and author contributions are available at Annals.org.

References

1. **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;9:1-12. [PMID: 22629234] doi:10.1371/journal.pmed.1001221
2. **Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al; PROGRESS Group.** Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Med. 2013;10: e1001381. [PMID: 23393430] doi:10.1371/journal.pmed.1001381
3. **Knottnerus JA.** Diagnostic prediction rules: principles, requirements and pitfalls. Prim Care. 1995;22:341-63. [PMID: 7617791]
4. **Lamain-de Ruiter M, Kwee A, Naaktgeboren CA, de Groot I, Evers IM, Groenendaal F, et al.** External validation of prognostic models to predict risk of gestational diabetes mellitus in one Dutch cohort: prospective multicentre cohort study. BMJ. 2016;354:i4338. [PMID: 27576867] doi:10.1136/bmj.i4338
5. **Riley RD, Hayden JA, Steyerberg EW, Moons KG, Abrams K, Kyzas PA, et al; PROGRESS Group.** Prognosis Research Strategy (PROGRESS) 2: prognostic factor research. PLoS Med. 2013;10: e1001380. [PMID: 23393429] doi:10.1371/journal.pmed.1001380
6. **Collins GS, Reitsma JB, Altman DG, Moons KG.** Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med. 2015;162:55-63. [PMID: 25560714] doi:10.7326/M14-0697
7. **Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al.** Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162:W1-73. [PMID: 25560730] doi:10.7326/M14-0698
8. **Collins GS, Mallett S, Omar O, Yu LM.** Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. BMC Med. 2011;9:103. [PMID: 21902820] doi:10.1186/1741-7015-9-103

9. **Kansagara D, Englander H, Salanitro A, Kagen D, Theobald C, Freeman M, et al.** Risk prediction models for hospital readmission: a systematic review. JAMA. 2011;306:1688-98. [PMID: 22009101] doi:10.1001/jama.2011.1515

10. **Steurer J, Haller C, Häuselmann H, Brunner F, Bachmann LM.** Clinical value of prognostic instruments to identify patients with an increased risk for osteoporotic fractures: systematic review. PLoS One. 2011;6:e19994. [PMID: 21625596] doi:10.1371/journal.pone.0019994

11. **Altman DG.** Prognostic models: a methodological framework and review of models for breast cancer. Cancer Invest. 2009;27:235-43. [PMID: 19291527] doi:10.1080/07357900802572110

12. **Shariat SF, Karakiewicz PI, Suardi N, Kattan MW.** Comparison of nomograms with other methods for predicting outcomes in prostate cancer: a critical analysis of the literature. Clin Cancer Res. 2008;14:4400-7. [PMID: 18628454] doi:10.1158/1078-0432.CCR-07-4713

13. **Counsell C, Dennis M.** Systematic review of prognostic models in patients with acute stroke. Cerebrovasc Dis. 2001;12:159-70. [PMID: 11641579]

14. **Perel P, Prieto-Merino D, Shakur H, Clayton T, Lecky F, Bouamra O, et al.** Predicting early death in patients with traumatic bleeding: development and validation of prognostic model. BMJ. 2012;345:e5166. [PMID: 22896030] doi:10.1136/bmj.e5166

15. **Damen JA, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, et al.** Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ. 2016;353:i2416. [PMID: 27184143] doi:10.1136/bmj.i2416

16. **Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Minhas R, Sheikh A, et al.** Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. BMJ. 2008;336:1475-82. [PMID: 18573856] doi:10.1136/bmj.39609.449676.25

17. **Graham R, Mancher M, Miller Wolman D, Greenfield S, Steinberg E, eds.** Clinical Practice Guidelines We Can Trust. Washington, DC: National Academies Pr; 2011.

18. **Goff DC Jr, Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, Gibbons R, et al; American College of Cardiology/American Heart Association Task Force on Practice Guidelines.** 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Circulation. 2014;129:S49-73. [PMID: 24222018] doi:10.1161/01.cir.0000437741.48606.98

19. **Rabar S, Lau R, O'Flynn N, Li L, Barry P; Guideline Development Group.** Risk assessment of fragility fractures: summary of NICE guidance. BMJ. 2012;345:e3698. [PMID: 22875946] doi:10.1136/bmj.e3698

20. **Centre for Reviews and Dissemination.** Systematic Reviews: CRD's Guidance for Undertaking Reviews in Health Care. York, United Kingdom: University of York; 2009.

21. **Higgins JPT, Green S, eds.** Cochrane Handbook for Systematic Reviews of Interventions. Chichester, United Kingdom: Wiley-Blackwell; 2011.

22. **Debray TP, Damen JA, Snell KI, Ensor J, Hooft L, Reitsma JB, et al.** A guide to systematic review and meta-analysis of prediction model performance. BMJ. 2017;356:i6460. [PMID: 28057641] doi:10.1136/bmj.i6460

23. **Hayden JA, van der Windt DA, Cartwright JL, Côté P, Bombardier C.** Assessing bias in studies of prognostic factors. Ann Intern Med. 2013;158:280-6. [PMID: 23420236] doi:10.7326/0003-4819-158-4-201302190-00009

24. **Higgins JPT, Savović J, Page MJ, Sterne JAC, ROB2 Development Group.** A revised tool for assessing risk of bias in randomized trials. In: Chandler J, McKenzie J, Boutron I, Welch V, eds. Cochrane Methods. London: Cochrane; 2018:1-69.

25. **Sterne JA, Hernán MA, Reeves BC, Savovic J, Berkman ND, Viswanathan M, et al.** ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016;355:i4919. [PMID: 27733354] doi:10.1136/bmj.i4919

26. **Moons KG, de Groot JA, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al.** Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med. 2014;11:e1001744. [PMID: 25314315] doi:10.1371/journal.pmed.1001744

27. **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:W1-W33. doi:10.7326/M18-1377

28. **Moher D, Schulz KF, Simera I, Altman DG.** Guidance for developers of health research reporting guidelines. PLoS Med. 2010;7:e1000217. [PMID: 20169112] doi:10.1371/journal.pmed.1000217

29. **Whiting P, Wolff R, Mallett S, Simera I, Savovic J.** A proposed framework for developing quality assessment tools. Syst Rev. 2017;6:204. [PMID: 29041953] doi:10.1186/s13643-017-0604-6

30. **Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al; QUADAS-2 Group.** QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155:529-36. [PMID: 22007046] doi:10.7326/0003-4819-155-8-201110180-00009

31. **Whiting P, Savovic J, Higgins JP, Caldwell DM, Reeves BC, Shea B, et al; ROBIS group.** ROBIS: a new tool to assess risk of bias in systematic reviews was developed. J Clin Epidemiol. 2016;69:225-34. [PMID: 26092286] doi:10.1016/j.jclinepi.2015.06.005

32. **Canet J, Gallart L, Gomar C, Paluzie G, Vallès J, Castillo J, et al; ARISCAT Group.** Prediction of postoperative pulmonary complications in a population-based surgical cohort. Anesthesiology. 2010;113:1338-50. [PMID: 21045639] doi:10.1097/ALN.0b013e3181fc6e0a

33. **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;66:268-77. [PMID: 23116690] doi:10.1016/j.jclinepi.2012.06.020

34. **Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG.** Prognosis and prognostic research: what, why, and how? BMJ. 2009;338:b375. [PMID: 19237405] doi:10.1136/bmj.b375

35. **Harrell FE.** Regression Modeling Strategies, With Applications to Linear Models, Logistic Regression, and Survival Analysis. New York: Springer; 2001.

36. **Hemingway H, Croft P, Perel P, Hayden JA, Abrams K, Timmis A, et al; PROGRESS Group.** Prognosis Research Strategy (PROGRESS) 1: a framework for researching clinical outcomes. BMJ. 2013;346:e5595. [PMID: 23386360] doi:10.1136/bmj.e5595

37. **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. [PMID: 20353578] doi:10.1186/1741-7015-8-20

38. **Altman DG, Vergouwe Y, Royston P, Moons KG.** Prognosis and prognostic research: validating a prognostic model. BMJ. 2009;338:b605. [PMID: 19477892] doi:10.1136/bmj.b605

39. **Moons KG, Altman DG, Vergouwe Y, Royston P.** Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ. 2009;338:b606. [PMID: 19502216] doi:10.1136/bmj.b606

40. **Royston P, Moons KG, Altman DG, Vergouwe Y.** Prognosis and prognostic research: developing a prognostic model. BMJ. 2009;338:b604. [PMID: 19336487] doi:10.1136/bmj.b604

41. **Knottnerus JA.** Between iatrotropic stimulus and interiatric referral: the domain of primary care research. J Clin Epidemiol. 2002;55:1201-6. [PMID: 12547450]

42. **Oudega R, Hoes AW, Moons KG.** The Wells rule does not adequately rule out deep venous thrombosis in primary care patients. Ann Intern Med. 2005;143:100-7. [PMID: 16027451]

Current Author Addresses: Drs. Wolff, Westwood, and Kleijnen: Kleijnen Systematic Reviews Ltd, Unit 6, Escrick Business Park, Riccall Road, Escrick, York YO19 6FD, United Kingdom.

Drs. Moons and Reitsma: Julius Centre for Health Sciences and Primary Care, UMC Utrecht, Utrecht University, PO Box 85500, 3508 GA Utrecht, the Netherlands.

Dr. Riley: Centre for Prognosis Research, Research Institute for Primary Care and Health Sciences, Keele University, Staffordshire ST5 5BG, United Kingdom.

Dr. Whiting: NIHR CLAHRC West, University Hospitals Bristol NHS Foundation Trust and School of Social and Community Medicine, University of Bristol, Bristol BS1 2NT, United Kingdom.

Dr. Collins: Centre for Statistics in Medicine, NDORMS, University of Oxford, Botnar Research Centre, Windmill Road, Oxford OX3 7LD, United Kingdom.

Dr. Mallett: Institute of Applied Health Sciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, United Kingdom.

Author Contributions: Conception and design: R.F. Wolff, K.G.M. Moons, R.D. Riley, P.F. Whiting, M. Westwood, G.S. Collins, J.B. Reitsma, J. Kleijnen, S. Mallett.

Analysis and interpretation of the data: R.F. Wolff, K.G.M. Moons, R.D. Riley, P.F. Whiting, M. Westwood, G.S. Collins, J.B. Reitsma, J. Kleijnen, S. Mallett.

Drafting of the article: R.F. Wolff, K.G.M. Moons, P.F. Whiting, M. Westwood, S. Mallett.

Critical revision of the article for important intellectual content: R.F. Wolff, K.G.M. Moons, R.D. Riley, P.F. Whiting, M. Westwood, G.S. Collins, J.B. Reitsma, J. Kleijnen, S. Mallett.

Final approval of the article: R.F. Wolff, K.G.M. Moons, R.D. Riley, P.F. Whiting, M. Westwood, G.S. Collins, J.B. Reitsma, J. Kleijnen, S. Mallett.

Statistical expertise: K.G.M. Moons, R.D. Riley, G.S. Collins, J.B. Reitsma, S. Mallett.

Obtaining of funding: K.G.M. Moons, R.D. Riley, P.F. Whiting, G.S. Collins, J.B. Reitsma, J. Kleijnen, S. Mallett.

Administrative, technical, or logistic support: R.F. Wolff, K.G.M. Moons, J. Kleijnen, S. Mallett.

Collection and assembly of data: R.F. Wolff, K.G.M. Moons, R.D. Riley, P.F. Whiting, M. Westwood, G.S. Collins, J.B. Reitsma, J. Kleijnen, S. Mallett.

Appendix: Members of the PROBAST Group

PROBAST Steering Group

Members of the PROBAST Group who authored this work: Robert F. Wolff, MD (Kleijnen Systematic Reviews, York, United Kingdom); Prof. Karel G.M. Moons, PhD (Julius Center for Health Sciences and Primary Care and Cochrane Netherlands, University Medical Center (UMC) Utrecht, Utrecht University, Utrecht, the Netherlands); Prof. Richard D. Riley, PhD (Keele University, Keele, United Kingdom); Penny F. Whiting, PhD (University Hospitals Bristol NHS Foundation Trust and University of Bristol, Bristol, United Kingdom); Marie Westwood, PhD (Kleijnen Systematic Reviews, York, United Kingdom); Prof. Gary S. Collins, PhD (Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, United Kingdom); Johannes B. Reitsma, MD, PhD (Julius Center for Health Sciences and Primary Care and Cochrane Netherlands, UMC Utrecht, Utrecht University, Utrecht, the Netherlands); Prof. Jos Kleijnen, MD, PhD (Kleijnen Systematic Reviews, York, United Kingdom, and School for Public Health and Primary Care, Maastricht University, Maastricht, the Netherlands); and Sue Mallett, DPhil (Institute of Applied Health Research, University of Birmingham, Birmingham, United Kingdom).

PROBAST Delphi Group

Members of the PROBAST group who were nonauthor contributors: Prof. Doug Altman, PhD (Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, United Kingdom); Prof. Patrick Bossuyt, PhD (Division of Clinical Methods & Public Health, University of Amsterdam, Amsterdam, the Netherlands); Prof. Nancy R. Cook, ScD (Brigham and Women's Hospital, Boston, Massachusetts); Gennaro D'Amico, MD (Ospedale Vincenzo Cervello, Palermo, Italy); Thomas P.A. Debray, PhD, MSc (Julius Center for Health Sciences and Primary Care and Cochrane Netherlands, UMC Utrecht, Utrecht University, Utrecht, the Netherlands); Prof. Jon Deeks, PhD (Institute of Applied Health Research, University of Birmingham, Birmingham, United Kingdom); Joris de Groot, PhD (Philips Image Guided Therapy Systems, Best, the Netherlands); Emanuele di Angelantonio, PhD, MSc (Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom); Prof. Tom Fahey, MD, MSc (Royal College of Surgeons in Ireland, Dublin, Ireland); Prof. Frank Harrell, PhD (Department of Biostatistics, Vanderbilt University, Nashville, Tennessee); Prof. Jill A. Hayden, PhD (Department of Community Health and Epidemiology, Dalhousie University, Halifax, Nova Scotia, Canada); Martijn W. Heymans, PhD (Department of Epidemiology and Biostatistics, Amsterdam Public Health Research Institute, Vrije Universiteit UMC, Amsterdam, the Netherlands); Lotty Hooft, PhD (Julius Center for Health Sciences and Primary Care and Cochrane Netherlands, UMC Utrecht, Utrecht University, Utrecht, the Netherlands); Prof. Chris Hyde, PhD (Institute of Health Research, University of Exeter Medical School, Exeter, United Kingdom); Prof. John Ioannidis, MD, DSc (Meta-Research Innovation Center at Stanford, Stanford University, Palo Alto, California); Prof. Alfonso Iorio, MD, PhD (Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada); Stephen Kaptoge, PhD (Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom); Prof. André Knottnerus, MD, PhD (Department of Family Medicine, Maastricht University, Maastricht, the Netherlands); Mariska Leeflang, PhD, DVM (Department of Clinical Epidemiology, Biostatistics and Bioinformatics, University of Amsterdam, Amsterdam, the Netherlands); Frances Nixon, BSc (National Institute for Health and Care Excellence, Manchester, United Kingdom); Prof. Pablo Perel, MD, PhD, MSc (Centre for Global Chronic Conditions, London School of Hygiene and Tropical Medicine, London, United Kingdom); Bob Phillips, PhD, MMedSci (Centre for Reviews and Dissemination, York, United Kingdom); Heike Raatz, MD, MSc (Kleijnen Systematic Reviews, York, United Kingdom); Rob Riemsma, PhD (Kleijnen Systematic Reviews, York, United Kingdom); Prof. Maroeska Rovers, PhD (Departments of Operating Rooms and Health Evidence, Radboud UMC, Nijmegen, the Netherlands); Anne W.S. Rutjes, PhD, MHSc (Institute of Social and Preventive Medicine and Institute of Primary Health Care, University of Bern, Bern, Switzerland); Prof. Willi Sauerbrei, PhD (Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany); Stefan Sauerland, MD, MPH (Institute for Quality and Efficiency in Healthcare, Cologne, Germany); Fülöp Scheibler, PhD, MA (UMC Schleswig-Holstein, Kiel, Germany); Prof. Rob Scholten, MD, PhD (Julius Center for Health Sciences and Primary Care and Cochrane Netherlands, UMC Utrecht, Utrecht University, Utrecht, the Netherlands); Ewoud Schuit, PhD, MSc (Julius Center for Health Sciences and Primary Care and Cochrane Netherlands, UMC Utrecht, Utrecht University, Utrecht, the Netherlands); Prof. Ewout Steyerberg, PhD (Department of Public Health, Erasmus UMC, Rotterdam, and Department of Biomedical Data Sciences, Leiden UMC, Leiden, the Netherlands); Toni Tan, MSc (National Institute for Health and Care Excellence, Manchester, United Kingdom); Gerben ter Riet, MD, PhD (Department of General Practice, University of Amsterdam, Amsterdam, the Netherlands); Prof. Danielle van der Windt, PhD (Centre for Prognosis Research, Keele University, Keele, United Kingdom); Yvonne Vergouwe, PhD (Department of Public Health, Erasmus UMC, Rotterdam, the Netherlands); Andrew Vickers, PhD (Memorial Sloan-Kettering Cancer Center, New York, New York); and Angela M. Wood, PhD (Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom).

The Delphi group members made substantial contributions to the conception and design, acquisition of data, or analysis and interpretation of the data; they drafted the article or revised it critically for important intellectual content; and they approved the final version to be published.